Università degli Studi di Padova
Dipartimento di Scienze Statistiche
Corso di Laurea Magistrale
in Scienze Statistiche

TESI DI LAUREA

INFERENZA NEL MODELLO DI POISSON PER DATI DI SOPRAVVIVENZA: UN CONFRONTO CON GLI STIMATORI NON DISTORTI IN MEDIANA

Relatore Prof.ssa Giuliana Cortese
Dipartimento di Scienze Statistiche

Correlatore Dott. Euloge Clovis Kenne Pagui
Dipartimento di Scienze Statistiche

Laureando Federico Baratin
Matricola 1134583

Anno Accademico 2018/2019

INDICE

INTRODUZIONE

L'analisi della sopravvivenza rappresenta un ramo importante nel mondo della statistica e fa capo a tutte quelle metodologie che studiano la distribuzione del tempo di comparsa di un evento. Il proprio campo di applicazione è assai vario e spazia da quello medico, il cui interesse potrebbe essere la sopravvivenza di un gruppo di pazienti sottoposti a una cura sperimentale, a quello ingegneristico, dove l'evento di interesse può essere rappresentato dal guasto di un apparecchio meccanico. Vi sono molteplici strumenti statistici che possono essere applicati in questo particolare tipo di contesto, uno di questi è il modello di Poisson, il quale dev'essere opportunamente adattato per i dati di sopravvivenza. Nel Capitolo 1 della tesi si procede a una breve presentazione del modello di Poisson, alla famiglia di modelli statistici a cui appartiene, ovvero i Modelli Lineari Generalizzati, e vengono riportati alcuni casi particolari del suo utilizzo. Dopo un breve excursus sulle principali definizioni e concetti riguardanti la sopravvivenza e sul modello a rischi proporzionali di Cox, il Capitolo 2 presenta il modello di Poisson per dati di sopravvivenza. Nonostante tale modello non sia stato specificatamente ideato per un contesto di analisi di sopravvivenza, in questo capitolo viene dimostrato come la verosimiglianza del modello di Poisson, sotto alcuni aggiustamenti ed ipotesi, risulti essere del tutto equivalente a quella per osservazioni censurate a destra di un modello esponenziale a rischi proporzionali costanti a tratti, giustificandone quindi l'utilizzo in questo campo. Il Capitolo 3 di questo lavoro è incentrato sulla teoria della verosimiglianza, in particolare viene presentato un nuovo stimatore ottenuto tramite una modifica della funzione punteggio, il quale gode della proprietà di non distorsione in mediana. La sua introduzione in questo lavoro è giustificata dall'obiettivo di confrontare, in una fase successiva, questo nuovo stimatore con quello "classico", ottenuto dalla funzione di verosimiglianza definita al Capitolo 2, e con lo stimatore che restituisce le stime dei coefficienti non distorte in media, presentato da Firth. Si continua con il Capitolo 4, nel quale è stata svolta un'analisi dei dati. Il dataset proviene dal Registro Tumori Norvegese e riporta una serie di informazioni riguardanti la mortalità di un gruppo di soggetti per i quali è stato diagnosticato un cancro al seno tra il 1965 e il 1974. Inizialmente vengono condotte alcune analisi

esplorative, al fine di derivare informazioni inerenti la sopravvivenza dei pazienti e successivamente viene applicato il modello di Poisson. Infine si confrontano i risultati ottenuti con i modelli che restituiscono le stime dei parametri non distorte in mediana e in media. Nel Capitolo 5 vengono svolte delle simulazioni Monte Carlo, tramite il software R, con lo scopo di confrontare le proprietà dello stimatore non distorto in mediana con lo stimatore "classico" e di Firth, per differenti scelte della numerosità campionaria e percentuale di censura. Per concludere, il Capitolo 6 si propone di approfondire brevemente un modello statistico alternativo a quello di Poisson presentato, ovvero il Poisson-Weibull, utilizzato anch'esso in contesti di analisi di sopravvivenza.

CAPITOLO I

REVISIONE DELLA LETTERATURA

1.1 Introduzione ai Modelli Lineari Generalizzati e al modello di Poisson

Il primo capitolo della tesi si propone di introdurre il modello di regressione di Poisson, con un breve excursus sulla famiglia di modelli statistici a cui esso appartiene: i Modelli Lineari Generalizzati.

Il modello di Poisson è ampiamente utilizzato nei contesti in cui la variabile risposta Y rappresenta un dato di conteggio o una frequenza.

Per arrivare a comprendere adeguatamente le caratteristiche intrinseche del modello in oggetto, ipotizziamo di osservare $y_1, \dots, y_n$ realizzazioni generate dalle variabili casuali $Y_1, \dots, Y_n$ indipendenti e identicamente distribuite secondo una legge Poisson, di media $\mu_1, \dots, \mu_n$.

La relativa funzione di densità viene espressa come:

$$f(y_i|\,\mu_i\,) = \frac{e^{-\mu_i}\,\mu_i{}^{y_i}}{y_i!}\,, con\; y_i \in N\; e\; \mu_i > 0. \qquad (1.1)$$

Il modello di Poisson appartiene a un'ampia classe di modelli statistici di regressione, definiti *Modelli Lineari Generalizzati* (GLM).

I Modelli Lineari Generalizzati costituiscono un'estensione del modello lineare di regressione e servono anch'essi allo studio della dipendenza in media di una variabile risposta da una o più variabili concomitanti. Un vantaggio di questa particolare tipologia di modelli, consiste nell'attenuazione di alcune ipotesi fondamentali nell'ambito del modello lineare, quali la linearità del modello di dipendenza, la normalità e l'omoschedasticità delle osservazioni. L'insieme dei modelli distributivi ammissibili per la variabile risposta viene ampliato, permettendo di includere molte delle distribuzioni di uso più comune, sia discrete che continue come, ad esempio, la distribuzione Normale, Gamma, Binomiale e

Poisson stessa. In particolare si assume che la variabile riposta Y appartenga alla famiglia di dispersione esponenziale.

La variabile aleatoria Y appartiene alla famiglia di dispersione esponenziale se la propria funzione di densità (o probabilità) è esprimibile nella forma:

$$f(y|\,\mu) = h(y)\, exp\{\psi(\mu)t(y) - k(\mu)\} \tag{1.2}$$

dove $h(\cdot)$ è una funzione dipendente dai dati, strettamente positiva, $\psi(\mu)$ è una funzione reale con dominio M, $t(y)$ è una statistica e $k(\mu)$ è una costante di normalizzazione finita, dipendente dai dati e da μ (Pace & Salvan, 2001).

È possibile dimostrare che il modello di Poisson appartiene alla famiglia di dispersione esponenziale. Si consideri la variabile aleatoria Y, distribuita secondo una legge Poisson di parametro μ, con funzione di densità specificata dalla (1.1):

$$f(y|\,\mu\,) = \frac{1}{y!} exp\{y \log \mu\, - \mu\,\} \tag{1.3}$$

con:

- $\psi(\mu) = \log \mu$, da cui si ricava: $\mu = e^{\psi(\mu)}$
- $t(y) = y$
- $h(y) = \frac{1}{y!}$
- $k(\mu) = \mu = e^{\psi(\mu)}$

Dalle definizioni precedenti è possibile ricavare la media e la varianza del processo Poisson:

- $E(Y) = k'(\mu) = e^{\psi(\mu)} = \mu$
- $Var(Y) = k''(\mu) = e^{\psi(\mu)} = \mu$

Nella classe dei Modelli Lineari Generalizzati si riconoscono tre specifiche componenti (Azzalini, 2001):

- una *componente casuale* Y_i, di media μ_i e dipendente da un parametro μ;

- una *componente sistematica* η_i, la quale entra in gioco attraverso un insieme di k variabili esplicative $x_1, \dots, x_k$, osservate su n individui e si suppone che siano in relazione con la variabile risposta. Tali variabili esplicative entrano nel modello per mezzo di un predittore lineare, ovvero una funzione delle stesse, lineare nei parametri $\beta_1, \dots, \beta_k$, definita come:

$$\eta_i = \sum_{j=1}^{k} x_{ij}\beta_j \qquad (1.4)$$

- una *funzione legame* g che mette in relazione il valor medio μ_i con il predittore lineare: $g(\mu_i) = \eta_i$.

La specificazione di un GLM viene completata attraverso la scelta di un'opportuna funzione legame, sulla quale non vengono imposte particolari restrizioni, se non che $g(\cdot)$ debba essere una funzione nota, monotona e derivabile. Per il modello Poisson viene scelto il legame canonico:

$$\eta_i = g(\mu_i) = \log \mu_i = \sum_{j=1}^{k} x_{ij}\beta_j \qquad (1.5)$$

Data la particolare scelta della funzione legame, si giunge alla famiglia dei *modelli Log-Lineari* (la cui applicazione verrà approfondita nel paragrafo successivo):

$$\log \mu_i = \sum_{j=1}^{k} x_{ij}\beta_j \qquad (1.6)$$

1.2 Casi particolari del modello di Poisson

1.2.1 Il modello Log-Lineare

Il modello Log-Lineare rappresenta uno strumento statistico, appartenente alla classe dei *GLM*, di analisi simmetrica per lo studio delle interazioni e delle associazioni tra le variabili esplicative. È particolarmente adatto quando si analizzano dati disposti in tabelle di contingenza, in cui le osservazioni sono classificate sulla base di un insieme di modalità di due o più variabili esplicative. Il caso non banale più elementare è rappresentato da un numero di classi pari a due: se si considera infatti una tabella a doppia entrata, le cui I righe e J colonne sono le modalità di due variabili qualitative A e B rispettivamente, per $i = 1, \dots, I$ e $j = 1, \dots, J$ e lo schema di campionamento non prevede una numerosità campionaria prefissata, allora le frequenze di cella Y_{ij} sono variabili indipendenti e identicamente distribuite secondo una legge Poisson di media μ_{ij} (Pace & Salvan, 2001). In questo contesto il modello Log-Lineare assume che il logaritmo delle frequenze di cella μ_{ij} sia pari a:

$$\log(\mu_{ij}) = \beta_0 + \beta_i^A + \beta_j^B + \beta_{ij}^{AB} \qquad (1.7)$$

Il parametro β_0 rappresenta l'effetto medio generale, stimato dalla numerosità complessiva. I parametri β_i^A e β_j^B sono gli effetti marginali che vengono stimati dalle frequenze marginali di A e B, μ_{i+} e μ_{+j}. Infine β_{ij}^{AB} è il parametro dell' interazione, la cui stima è ottenuta tramite le frequenze di cella μ_{ij}. Quest'ultimo parametro risulta essere di particolare interesse in molte tipologie di analisi come quelle in campo biomedico. Infatti, se A rappresenta un determinato fattore di rischio per la presenza/assenza di una particolare malattia B, allora il termine β_{ij}^{AB} rappresenta l'associazione tra queste due variabili e una sua significatività è rilevante ai fini dell'analisi. Il modello (1.7) è un modello saturo, ovvero contempla tutti i possibili effetti e riproduce sempre esattamente le frequenze nella tabella di contingenza. Tuttavia spesso si procede nell'identificare un modello più parsimonioso di quello saturo che spieghi altrettanto bene la distribuzione dei dati della tabella e che allo stesso tempo permetta un'interpretazione agevole dei parametri (Agresti, 2003).

1.2.2 Il modello moltiplicativo per la funzione di intensità

In alcuni contesti di studio si è direttamente interessati nel conteggio del numero di eventi che si verificano in un preciso intervallo temporale e, molto spesso, si vuole identificare il tasso con cui questi eventi si verificano. Il tasso descrive il rischio istantaneo del verificarsi di un evento in un particolare istante temporale; per essere più chiari, la probabilità di osservare un evento esattamente al tempo t in un intervallo temporale che intercorre tra t e $t+h$, diviso per la lunghezza dello stesso h, tende a una valore $\lambda(t)$ (per h che tende verso lo 0). In altri termini:

$$\lim_{h \to 0} \frac{\Pr\left(t \leq T < t+h \mid T \geq t \right)}{h} = \lambda(t), \qquad (1.8)$$

dove la funzione del tempo t, $\lambda(t)$, viene chiamata *tasso* o *funzione di intensità*.

In questo contesto, il processo di Poisson assume che i tempi di attesa tra due eventi successivi possano essere considerati indipendenti e identicamente distribuiti secondo una legge Esponenziale, con media pari a $\frac{1}{\lambda}$. In questo caso, la *funzione di intensità* rimane costante nel tempo, ovvero $\lambda(t) \equiv \lambda$. Il numero di eventi che si verificano fino al tempo t, indicato con $Y(t)$, si distribuisce secondo una legge Poisson, di parametro λt:

$$f(y_i | \lambda t) = \frac{e^{-\lambda t} \lambda t^{y_i}}{y_i!}, \text{con } \lambda t = \mu_t. \qquad (1.9)$$

Quindi il valor medio $E\big(Y(t)\big) = \lambda t = \mu_t$ rappresenta il numero medio di eventi nell'intervallo $[0, t]$.

Di conseguenza, si può procedere con un approccio che si basa sulla regressione di Poisson, dove il tasso dipende dalle variabili concomitanti x (Seeber, 2014):

$$\log \lambda = \log \frac{\mu_t}{t} = \alpha + \beta x \qquad (1.10)$$

La quantità precedente può essere riscritta come:

$$\log \mu_t = \alpha + \beta x + \log t, \qquad (1.11)$$

con $\log t$ inteso come offset. Passando alla forma esponenziale della (1.10), si ottiene un modello moltiplicativo per la *funzione di intensità*, ossia:

$$\lambda = e^{\alpha} e^{\beta x}, \qquad (1.12)$$

dove $\lambda_0 = e^{\alpha}$ viene definito *funzione hazard baseline*, mentre e^{β} è il fattore di proporzionalità, anche chiamato *rischio relativo*.

Per avere una più chiara comprensione del funzionamento del modello, si procede con un semplice esempio, i cui dati sono riportati nella Tabella 1.1.

Dimensione	**Ricorrenze**	**Tempo sotto osservazione**
≤ 3 cm	1	2, 3, 6, 8, 9, 10, 11, 13, 14, 16, 21, 22, 24, 26, 27
	2	7, 13, 15, 18, 23
	3	20
	4	24
> 3 cm	1	1, 5, 17, 18, 25
	2	18, 25
	3	4
	4	19

Tabella 1.1: Numero di tumori secondari comparsi dopo l'eradicazione del tumore primitivo per 31 pazienti maschi, con i rispettivi tempi di osservazione (in mesi), divisi per diametro tumorale.

Si dispone di un campione costituito da 31 pazienti maschi, affetti da cancro alla vescica. Viene registrato il numero di tumori secondari comparsi dopo un certo intervallo di tempo dall'eradicazione del tumore primitivo. Si definisce una variabile aleatoria X, che assume valore 1 se il tumore primitivo presenta un diametro maggiore di 3 cm, 0 altrimenti (minore o uguale a 3 cm).

Si consideri il modello di Poisson con tasso specificato dalla (1.12). Date le stime dei parametri α e β, pari a -1.95 e 0.385, rispettivamente, si può ottenere la stima per il tasso di ricorrenze, relativo ai soggetti con tumori di dimensioni minori o uguali a 3 cm: $e^{\alpha} = 0.142$, mentre il tasso per i tumori maggiori di 3 cm è circa 1.47 volte più alto. Se si vuole, infine, esprimere i risultati in termini di tempo di attesa tra l'evento di eradicazione del tumore primitivo e la comparsa di una neoplasia secondaria, si ottengono le medie stimate, sostituendo a $\frac{1}{\lambda}$ i tassi trovati in precedenza, pari a 7.06 mesi per i tumori più piccoli contro 4.80 mesi per i tumori più grandi, dimostrando un'occorrenza maggiore di seconda neoplasia per i tumori con dimensioni superiori ai 3 cm.

Si possono ottenere diverse specificazioni del modello presentato, a seconda che il processo sia caratterizzato da un tasso proporzionale o da un tasso che varia nel tempo, meglio definito come processo di Poisson non omogeneo. In questo lavoro non verranno approfonditi in quanto ci si focalizzerà sul modello di Poisson applicato nel contesto dell'analisi della sopravvivenza, introdotto nel capitolo che segue.

CAPITOLO II

IL MODELLO DI POISSON PER DATI DI SOPRAVVIVENZA

2.1 Introduzione ai modelli di sopravvivenza

L'analisi della sopravvivenza è una parte rilevante della statistica inferenziale e la sua peculiarità consiste nel mettere in rapporto un certo esito o "evento di interesse" con il fattore tempo. Tale modalità di analisi si riferisce all'insieme di metodi statistici che studiano la distribuzione del tempo di comparsa di un evento.

Il termine "tempo di sopravvivenza" viene usato in senso estensivo perché si applica anche a eventi che differiscono dalla morte. L'analisi di sopravvivenza riguarda, infatti, tutti quegli studi in cui si vuole analizzare l'incidenza di un determinato evento in un certo intervallo di tempo e prevede l'utilizzo di dati provenienti da studi di coorte o di follow-up. Il tempo di sopravvivenza assume significati differenti in relazione al tipo di evento a cui è interessato il ricercatore. L'esempio più comune riguarda il tempo che intercorre tra l'inizio dello studio e la morte del soggetto, oppure, riferendosi ad altri contesti come quello epidemiologico, il tempo di insorgenza di una patologia o il tempo di risposta a un trattamento a cui il paziente è sottoposto.

Quando si ha a che fare con questa particolare tipologia di modelli statistici, si deve tener conto di alcune loro particolari caratteristiche. In primo luogo la variabile dipendente o risposta rappresenta il tempo di attesa al verificarsi di un preciso evento; le osservazioni sono solitamente censurate a destra, nel senso che per alcuni soggetti può non verificarsi l'evento di interesse nel tempo in cui sono in studio; infine sono presenti variabili esplicative di cui si vuole valutare l'effetto sul tempo di attesa.

Verranno di seguito presentati alcuni concetti e definizioni che risultano essere necessari per proseguire lo studio, considerando i lavori di (Rodriguez, 2007) e (Allison, 1995).

2.2 Definizioni principali e meccanismi di censura

La durata del tempo di sopravvivenza o meglio dell' "episodio" di un soggetto è la realizzazione di una variabile aleatoria continua e non negativa T, caratterizzata da una specifica funzione di densità di probabilità $f(t)$ e da una funzione di ripartizione $F(t)$.

La funzione $F(t)$ rappresenta la probabilità che il soggetto abbia sperimentato l'evento di interesse entro il tempo t, in altri termini:

$$F(t) = \Pr(T < t) = \int_0^t f(x)dx \qquad (2.1)$$

In questi ambiti, tuttavia, si è soliti lavorare con la *funzione di sopravvivenza*, la quale esprime la probabilità che il soggetto non abbia sperimentato l'evento di interesse fino al tempo t. Tale funzione viene espressa come complemento a uno della funzione di ripartizione,

$$S(t) = \Pr(T \geq t) = 1 - F(t) = \int_t^\infty f(x)dx \qquad (2.2)$$

Le funzioni di sopravvivenza e di densità sono due concetti matematicamente equivalenti. Tuttavia, nell'analisi della sopravvivenza, si preferisce ricorrere alla seconda delle due funzioni presentate, dato che permette una descrizione più intuitiva del fenomeno oggetto di studio (Blossfeld, Golsch, & Rohwer, 2012). La funzione di sopravvivenza $S(t)$ presenta un comportamento del tutto particolare, osservabile all'aumentare del tempo t: all'inizio del periodo di tempo considerato, cioè per un valore di t pari a 0, la funzione assume valore 1, poiché nessun soggetto in studio ha ancora sperimentato l'evento di interesse e quindi tutti sono considerati come "sopravvissuti" all'evento. Col passare del tempo, quando si verificano gli eventi, la funzione di sopravvivenza decresce verso lo 0. Non è detto, tuttavia, che tutti gli individui considerati sperimenteranno l'evento,

pertanto, quando lo studio terminerà, potrebbe esserci una certa proporzione di sopravvissuti (per esempio, se l'evento di interesse è il matrimonio, potrebbero esserci persone che non si sposeranno mai) e il valore della funzione di sopravvivenza rimarrà stabile al suo livello precedente.

Una caratterizzazione alternativa della distribuzione di T è data dalla funzione di rischio o tasso di rischio istantaneo. In formule, si ha:

$$h(t) = \lim_{dt \to 0} \frac{\Pr(t \leq T < t + dt | \, T \geq t)}{dt} \tag{2.3}$$

dove $h(t)$ esprime il rischio che l'evento di interesse accada nell'intervallo infinitesimo dt condizionatamente al fatto che non si è verificato fino al tempo t. Si può esprimere la funzione di rischio come rapporto tra la funzione di densità di probabilità $f(t)$ e la funzione di sopravvivenza $S(t)$:

$$h(t) = \frac{f(t)}{S(t)} \tag{2.4}$$

Dalla (2.2) si osserva che $-f(t)$ è la derivata di $S(t)$. Di conseguenza, si può riscrivere l'espressione (2.4) come:

$$h(t) = -\frac{d}{dt} \log\big(S(t)\big) \tag{2.5}$$

Procediamo a integrare la quantità ricavata al (2.5), ottenendo così una formulazione per la probabilità di sopravvivere fino al tempo t in funzione del rischio di tutte le occorrenze sino a t:

$$S(t) = exp\left\{-\int_0^t h(x)dx\right\} \tag{2.6}$$

Si può arrivare alla definizione di rischio integrato, la quale rappresenta l'integrale della funzione di rischio fino al tempo t, ovvero alla somma dei rischi nell'intervallo di tempo $[0, t]$.

$$H(t) = \int_0^t h(x)dx = -\ln\big(S(t)\big) \tag{2.7}$$

L'utilità di questa funzione risiede nella possibilità, passando per $S(t)$, di ricavare informazioni per via indiretta sulla funzione di rischio, laddove non si riesce a stimarla direttamente.

Questi risultati mostrano come la funzione di sopravvivenza e la funzione di rischio siano caratterizzazioni alternative, ma allo stesso tempo equivalenti, della distribuzione di T. Data la funzione di sopravvivenza si può differenziare per ottenere la funzione di densità di probabilità e, successivamente, la funzione di rischio, usando la (2.5). Al contrario, data la funzione di rischio si può integrare per ricavare la funzione di rischio integrato e poi ricorrere all'esponenziale per ottenere la funzione di sopravvivenza, usando la (2.6).

Negli studi longitudinali si ha a che fare con soggetti i cui tempi di osservazione sono censurati a destra e quindi risultano essere incompleti. Spesso ciò è dovuto al fatto che tali soggetti non hanno sperimentato l'evento di interesse fino al termine dello studio, oppure sono entrati in osservazione ma sono successivamente sfuggiti al controllo dello sperimentatore.

Vi sono diversi meccanismi di censura che possono essere applicati nello studio (Allison, 1995):

- **Tipo I**: viene fissato il tempo di censura t (che è sotto il controllo dello sperimentatore) in cui vengono seguite n unità statistiche e si ipotizza che il numero di unità che sperimenteranno l'evento di interesse sia casuale. Verranno considerate come censure tutti quei soggetti che non avranno subito l'evento entro il tempo t. I tempi di censura possono variare tra i diversi individui che partecipano allo studio;
- **Tipo II**: un campione di n di unità statistiche viene seguito fino a quando d soggetti sperimenteranno l'evento di interesse. Pertanto, tutti gli individui che non avranno sperimentato l'evento dopo il d-esimo successo, verranno censurati;
- **Censura casuale**: a ogni individuo viene associato un tempo di censura e di durata potenziale, rappresentati da due variabili casuali indipendenti, C_i e T_i rispettivamente. Per ogni soggetto, quindi, osserveremo il minimo tra questi due tempi e un indicatore d_i, che informerà lo sperimentatore se l'osservazione è terminata per il verificarsi dell'evento o per la censura del soggetto;

Nella censura di tipo I e casuale, assumiamo che il meccanismo di censura sia non informativo. Questo rappresenta un'assunzione che caratterizza la maggior parte dei metodi per l'analisi dei dati di durata, in quanto non viene tenuto conto dell'informazione prodotta dalle censure, assumendo che il processo generatore della censura stessa sia indipendente dall'evento di interesse. Ne consegue che la distribuzione dei valori censurati non dipenderà dai parametri di interesse del modello in uso. Nel caso delle censure informative come la censura di tipo II, invece, il meccanismo di censura seleziona soggetti con durate specifiche rispetto al fenomeno in studio e l'ipotesi di indipendenza tra tempo di censura e il momento in cui si verifica l'evento di interesse non è valida. In questo secondo caso, si giungerà a una distorsione delle stime dei parametri e a un'inferenza statistica sulla sopravvivenza inadeguata.

2.3 Modello a rischi proporzionali

Il cuore del capitolo risiede nel definire il modello di Poisson applicato in un contesto diverso da quello in cui si è abituati a vederlo: l'analisi della sopravvivenza. In questo particolare ambito sono stati proposti alcuni strumenti, tra cui il modello semi-parametrico a rischi proporzionali introdotto da (Cox, 1972). Non ci si soffermerà eccessivamente nella sua descrizione, ma ne verranno riportate le principali caratteristiche, per comprendere meglio le differenze e gli aggiustamenti necessari per arrivare a specificare il modello di Poisson.

La funzione di rischio viene modellata attraverso una componente non parametrica $\lambda_0(t)$, definita come *baseline* o *rischio del gruppo di base*, che descrive il rischio per gli individui con caratteristiche $x_i = 0$, e da una componente parametrica $\exp(x_i'\beta)$, ovvero il *rischio relativo*. Di conseguenza, il rischio al tempo t per uno specifico individuo i avente caratteristiche x_i è rappresentato da:

$$\lambda_i(t|x_i) = \lambda_0(t)\exp(x_i'\beta). \qquad (2.8)$$

L'ipotesi sottostante il modello è la proporzionalità dei rischi, ovvero il rapporto dei rischi di due individui con un diverso set di variabili esplicative è costante al variare del tempo. Si tratta di un modello robusto, in quanto i coefficienti di regressione approssimano bene i risultati di un corretto modello parametrico e fornisce informazioni primarie sulle funzioni di sopravvivenza e di rischio di diversi gruppi di individui con un numero minimo di assunzioni.

Il modello specificato è dotato di grande flessibilità, in quanto può incorporare covariate tempo-dipendenti, ovvero variabili che cambiano nel corso del tempo, ed essere eventualmente esteso a effetti non proporzionali. Queste due estensioni possono essere facilmente combinate tramite semplici modifiche dell'espressione (2.8).

Dalla letteratura emergono diversi approcci per la stima del modello illustrato in precedenza. È noto che il metodo principe per la stima di un modello statistico consiste nella massimizzazione della relativa funzione di log-verosimiglianza.

Tuttavia, dato che ci si trova di fronte a un modello semi-parametrico, non è possibile identificare un unico set di parametri che massimizzi la verosimiglianza, poiché sono ignote le distribuzioni dei tempi di sopravvivenza. (Cox, 1972) nel proprio lavoro propone di massimizzare la funzione di verosimiglianza parziale (attraverso metodi numerici iterativi), derivando in questo modo le stime dei parametri di interesse dopo aver eliminato il parametro di disturbo $\lambda_0(t)$. Tali stime risulteranno essere asintoticamente normali e non distorte.

Una strada alternativa e preferibile alla precedente prevede un approccio parametrico, che consiste in uno strumento flessibile, di ampio utilizzo e strettamente connesso alla regressione di Poisson, di diretto interesse per lo studio. Se si procede a suddividere l'asse del tempo in J intervalli e si assume che il rischio di base rimanga costante all'interno di ogni intervallo, allora si ha $\lambda_0(t) = \lambda_j$, giungendo a definire per il soggetto $i-esimo$ un modello esponenziale a rischi proporzionali e costanti a tratti del tipo:

$$\lambda_{ij} = \lambda_j \exp(x_i'\beta), \tag{2.9}$$

dove λ_j rappresenta il rischio di base riferito all'intervallo j e $x_i'\beta$ il rischio relativo per un individuo avente caratteristiche x_i, in riferimento al rischio del gruppo di base, a qualunque tempo t.

Applicandovi una trasformazione logaritmica, si ottiene il modello seguente:

$$log\lambda_{ij} = \alpha_j + x_i'\beta, \tag{2.10}$$

dove $\alpha_j = log\lambda_j$ è il logaritmo del rischio di base. Questo modello rappresenta il punto di partenza per ciò che si andrà a definire nel paragrafo seguente e allo stesso tempo il fulcro del capitolo: il modello Poisson per dati di sopravvivenza.

2.4 Modello di Poisson per dati di sopravvivenza

Il modello di Poisson può essere applicato anche in un contesto che differisce da quelli specificati nel capitolo precedente: l'analisi della sopravvivenza. Autori come (Holford, 1980), (Laird & Olivier, 1981), in alcuni dei lavori da loro prodotti, hanno evidenziato come il modello esponenziale a rischi proporzionali costanti a tratti risulti essere equivalente a uno specifico modello di Poisson. Prima di passare agli aspetti formali di tale modello, occorre specificare alcuni elementi che risulteranno essere utili nel capire fino in fondo la metodologia presentata nel paragrafo, tratti dal lavoro di (Rodriguez, 2007).

Inizialmente si necessita di definire il periodo di esposizione per un soggetto i, che verrà denominato come t_i. Viene definito inoltre come t_{ij} il tempo in cui l'individuo i-esimo è a rischio di subire l'evento di interesse nell'intervallo j, delimitato dagli estremi $[\tau_{j-1}, \tau_j]$, con $j = 1, \dots, J$ e $\tau_0 = 0$. Nel caso in cui il soggetto non dovesse sperimentare l'evento entro questo intervallo, allora il tempo trascorso nell'intervallo eguaglierà l'ampiezza dello stesso, cioè $t_{ij} = \tau_j - \tau_{j-1}$. Se invece l'individuo subisce l'evento o viene censurato, allora il tempo trascorso nell'intervallo sarà pari alla differenza tra il tempo in cui l'individuo è stato esposto e l'estremo inferiore dell'intervallo, in altri termini: $t_{ij} = t_i - \tau_{j-1}$. Siccome si procederà a considerare solamente gli intervalli visitati dai soggetti, se l'individuo i ha sperimentato l'evento di interesse prima dell'inizio dell'intervallo j, e quindi $t_i < \tau_{j-1}$, allora il tempo trascorso in tutti gli intervalli successivi a questo sarà pari a 0.

Il secondo ingrediente necessario è un indicatore di evento o censura per il soggetto i-esimo, definito come d_i. Sia d_{ij}, l'indicatore che assumerà valore 1 o 0 a seconda che il soggetto i-esimo sperimenti l'evento di interesse nell'intervallo j o meno. Infine chiameremo con $j(i)$ l'intervallo dove t_i cade, ovvero l'intervallo j in cui il soggetto i sperimenta l'evento o viene censurato. Infatti, se t_i cade nell'intervallo $j(i)$, allora si avrà $d_{i1} = d_{i2} = \dots = d_{ij-2} = d_{ij-1} = 0$, mentre $d_{ij} = d_i$ per $j = j(i)$, cioè per l'ultimo intervallo visitato dall'individuo i-esimo.

Sulla base di quanto appena definito, è possibile stimare un modello esponenziale a rischi proporzionali costanti a tratti, trattando le quantità d_{ij} come se fossero delle osservazioni indipendenti provenienti da una distribuzione Poisson di media:

$$\mu_{ij} = t_{ij}\lambda_{ij}. \tag{2.11}$$

Il primo termine rappresenta, come precedentemente illustrato, il periodo di esposizione, per l'individuo i, nell'intervallo j mentre il secondo è il rischio del soggetto i-esimo di sperimentare l'evento di interesse nell'intervallo j.

2.4.1 Modello a tasso costante senza variabili esplicative

Prima di arrivare ad approfondire il modello che è di nostro diretto interesse, si presentano qui di seguito due modelli più semplici.

Il primo corrisponde al contesto in cui non si abbiano a disposizione variabili esplicative e il cui rischio, esprimibile come $\lambda_{ij} = \frac{\mu_{ij}}{t_{ij}}$, sia costante nel tempo t e fissato pari a λ_0 per tutti i soggetti. Partendo dall'espressione (2.11) si ottiene:

$$\mu_{ij} = t_{ij}\lambda_{ij} = t_{ij}\lambda_0, \quad \text{dove } \lambda_{ij} = \lambda_0 \qquad \forall\, i = 1, \dots, n \text{ e } \forall\, j = 1, \dots, J.$$

Applicando la trasformazione logaritmica alla precedente definizione, è possibile giungere alla formulazione del primo modello considerato, ovvero:

$$log\mu_{ij} = logt_{ij} + log\lambda_0. \tag{2.12}$$

2.4.2 Modello a tasso costante con variabili esplicative

Prendendo in considerazione il secondo caso, in cui sono presenti delle variabili esplicative che sono di diretto interesse per l'analisi, il modello visto in precedenza può essere modificato, esprimendo λ_{ij} come $e^{x_i'\tilde{\beta}} = e^{\beta_0 + x_i'\beta}$, dove e^{β_0} rappresenta il rischio per il gruppo di base e può essere posto pari a λ_0, mentre $e^{x_i'\beta}$ è il rischio relativo degli individui aventi caratteristiche i. Nella notazione precedente, il termine $\tilde{\beta}$ indica il vettore dei parametri comprendente l'intercetta β_0. Quindi, partendo dalla medesima relazione utilizzata in precedenza (2.11) e apportando le sostituzioni appena definite, è possibile giungere al seguente modello:

$$\mu_{ij} = t_{ij}\lambda_{ij} = t_{ij}\, e^{\beta_0 + x_i'\beta} = t_{ij}\, e^{\beta_0}\, e^{x_i'\beta} = t_{ij}\, \lambda_0 e^{x_i'\beta}$$

$$log\mu_{ij} = logt_{ij} + \beta_0 + x_i'\beta, \qquad \text{con } \beta_0 = log\lambda_0 \qquad \forall\, i = 1, \dots, n \qquad \forall j = 1, \dots, J \qquad (2.13)$$

2.4.3 Modello di regressione con tasso costante a tratti

Vi è infine un terzo caso, riguardante il modello che è di diretto interesse in questo lavoro. A differenza del primo modello visto in precedenza (2.12), non solo si assume la presenza di un insieme di covariate, ma anche che il rischio non sia più costante al variare del tempo t.

Per mettere in evidenza questa modifica, il rischio $\lambda_{ij} \equiv \lambda_{ij}(t)$ è posto pari a $e^{x_i'\tilde{\beta}(t)} = e^{\beta_0(t) + x_i'\beta}$, dove $e^{\beta_0(t)} = \lambda_0(t)$, il quale rappresenta il rischio per il gruppo di riferimento che varia nel tempo t, mentre il secondo termine $e^{x_i'\beta}$ è il rischio relativo dei soggetti con covariate i.

Nell'ambito del modello di Poisson applicato nel campo della sopravvivenza, l'asse temporale viene suddiviso in J intervalli, delimitati dagli estremi $[\tau_{j-1}, \tau_j]$,

con $j = 1, \dots, J$ e $\tau_0 = 0$, i quali non devono avere necessariamente le medesima ampiezza. All'interno di ciascun intervallo il rischio risulta essere costante, ma varia da un intervallo all'altro. Si assume infine che $\beta_0(t) = \sum_{j=1}^{J} I_j \beta_{0j}$ ed è quindi una funzione del tempo costante a tratti, con salti agli estremi degli intervalli. Il rischio viene assunto pari a $\lambda_{ij}(t) = \lambda_0(t) e^{x_i' \beta}$.

Pertanto $\lambda_0(t)$ può essere interpretato come la somma dei rischi di ciascun intervallo, ovvero: $\lambda_0(t) = e^{\beta_0(t)} = e^{\sum_{j=1}^{J} I_j \beta_{0j}} = \prod_{j=1}^{J} e^{I_j \beta_{0j}}$, dove I_j rappresenta l'indicatore dell'intervallo j-esimo che vale 1 se si fa riferimento all'intervallo j e 0 altrimenti; ogni β_{0j} è il rischio relativo all'intervallo j, che è quindi costante in questo stesso intervallo.

Partendo dalla relazione (2.11), è possibile giungere al modello desiderato:

$$\mu_{ij} = t_{ij} \lambda_{ij}(t) = t_{ij} e^{\beta_0(t)} e^{x_i' \beta} = t_{ij} e^{\sum_{j=1}^{J} I_j \beta_{0j}} e^{x_i' \beta}$$

Applicando la trasformazione logaritmica all'espressione precedente si ottiene:

$$log \mu_{ij} = log t_{ij} + x_i' \beta + \sum_{j=1}^{J} I_j \beta_{0j}. \qquad (2.14)$$

Il termine $I_j \beta_{0j}$ rappresenta il rischio relativo di base nell'intervallo j-esimo, preso su scala logaritmica, che risulta essere costante. Sotto l'ipotesi di rischi proporzionali, è possibile osservare come il modello esponenziale a rischi proporzionali costanti a tratti (riportato nella (2.10)) sia equivalente a un modello log-lineare di Poisson per pseudo-osservazioni (t_{ij}, d_{ij}), una per ogni combinazione di individui e intervalli. Il termine d_{ij} gioca il ruolo di variabile

risposta, il logaritmo del tempo di esposizione entra nel modello come offset e il termine $\alpha_j = \sum_{j=1}^{J} I_j \beta_{0j}$ rappresenta il rischio relativo costante nell'intervallo j.

Il termine μ_{ij} della (2.14) rappresenta il numero di eventi attesi per il soggetto i-esimo, che si verificano nell'intervallo j, avente estremi $[\tau_{j-1}, \tau_j]$. Si può assumere che tale valore atteso sia quello di una distribuzione Poisson per il numero di eventi osservati d_{ij}. Nell'ambito dell'analisi di sopravvivenza di cui si sta trattando in questo lavoro, ogni soggetto può sperimentare un solo evento di interesse (o in alternativa essere censurato) nel corso dello studio e questo si verifica in un determinato intervallo j, che rappresenta quindi l'intervallo in cui cade il soggetto stesso. Di conseguenza d_{ij} sarà pari a 1 se l'evento, per un soggetto i, si realizza in un determinato intervallo j e 0 altrimenti.

In precedenza è stato detto che la variabile d_{ij} è caratterizzata da una distribuzione Poisson di media $\mu_{ij} = t_{ij}\lambda_{ij}$. La funzione di densità del modello di Poisson per dati di sopravvivenza può essere quindi espressa come:

$$d_{ij} \sim Poisson(\mu_{ij}), \qquad \forall\, i = 1, \dots, n \qquad \forall\, j = 1, \dots, J$$

$$f(d_{ij} | \, \mu_{ij}) = \frac{e^{-\mu_{ij}} \, {\mu_{ij}}^{d_{ij}}}{d_{ij}!} \text{, con } d_{ij} \in \{0,1\} \text{ e } \mu_{ij} > 0. \qquad (2.15)$$

Come per il modello Poisson presentato al Capitolo 1, la funzione di densità appena specificata appartiene alla famiglia di dispersione esponenziale, esprimibile nella seguente forma:

$$f(d_{ij} | \, \mu_{ij}) = \frac{1}{d_{ij}!} exp\{d_{ij} \log \mu_{ij} - \mu_{ij}\}$$

con:

- $\psi(\mu_{ij}) = \log \mu_{ij}$, da cui si ricava: $\mu_{ij} = e^{\psi(\mu_{ij})}$
- $t(y) = d_{ij}$
- $h(y) = \frac{1}{d_{ij}!}$
- $k(\mu_{ij}) = \mu_{ij} = e^{\psi(\mu_{ij})}$ (2.16)

Se si indica con $\mu_{ij} = t_{ij}\lambda_{ij} = t_{ij}e^{\beta_0(t)+x_i'\beta}$, la funzione legame $g(\cdot)$ che mette in relazione il predittore lineare con la media del processo μ_{ij} è rappresentata da $g(\cdot) = \log(\cdot)$, che è il legame canonico per la distribuzione di Poisson. Sulla base di quanto definito, è possibile riscrivere la relazione come segue:

$$g\left(E(d_{ij})\right) = g(\mu_{ij}) = \log(\mu_{ij}) = log t_{ij} + \beta_0(t) + x_i'\beta \quad (2.17)$$

I diversi soggetti che fanno parte del campione osservato sono indipendenti tra loro. In quest' analisi della sopravvivenza si dispone di diverse osservazioni appartenenti allo stesso soggetto i, ciascuna relativa a ogni specifico intervallo j visitato dal soggetto stesso. Per esempio, se un individuo viene osservato per ventiquattro anni (quindi nel ventiquattresimo anno di osservazione sperimenta l'evento di interesse o viene censurato, uscendo così dallo studio) e i J intervalli in cui viene suddiviso l'asse temporale sono di ampiezza quinquennale, allora vi saranno cinque osservazioni per quel determinato individuo. Le prime quattro sono caratterizzate da un periodo di esposizione t_{ij} per il soggetto i-esimo di durata pari a cinque anni e, per ognuna di esse, l'indicatore d_{ij} sarà pari a 0, indicando che non è avvenuto l'evento per l'individuo stesso. L'ultima osservazione, invece, avrà una durata di quattro anni e l'indicatore d_{ij} assumerà valore unitario se il soggetto sperimenterà l'evento di interesse o valore nullo se verrà censurato. Per quanto riguarda queste diverse osservazioni dell'individuo i-esimo, invece, l'ipotesi di indipendenza potrebbe non essere così scontata. Infatti a primo impatto potrebbero non essere indipendenti, in quanto un soggetto che

sperimenta l'evento di interesse in un determinato intervallo j, non lo ha sicuramente sperimentato negli intervalli a questo precedenti, anche perché, come è stato detto, si lavora in un contesto in cui si può verificare un solo evento per ciascun individuo i.

Tuttavia, nel contesto del processo di Poisson, si assume che i tempi di attesa che intercorrono tra eventi successivi siano indipendenti e identicamente distribuiti secondo un processo Esponenziale, la cui media risulta essere pari a $\frac{1}{\lambda_{ij}}$. La distribuzione esponenziale gode della proprietà dell'assenza di memoria (Azzalini, 2001); infatti, supposto che un evento non si sia verificato nell'intervallo $[\tau_{j-1}, \tau_j]$, allora la probabilità che non si verifichi nell'intervallo immediatamente successivo, indicato come $[\tau_j, \tau_{j+1}]$, è la stessa dell'intervallo precedente. Per questo motivo anche le diverse osservazioni relative allo stesso individuo possono essere considerate indipendenti.

Considerando quindi di disporre di un campione indipendente per $i = 1, \dots, n$ e $j = 1, \dots, J$ e di dimensione $(n\, x\, J)$, la verosimiglianza del modello può essere espressa nella seguente forma:

$$L(\mu) = \prod_{i=1}^{n} \prod_{j=1}^{J} f(d_{ij} | \, \mu_{ij} \,)$$

$$L(\mu) = \prod_{i=1}^{n} \prod_{j=1}^{J} \frac{1}{d_{ij}!} exp\{d_{ij} \log \mu_{ij} \, - \mu_{ij} \} \propto \prod_{i=1}^{n} \prod_{j=1}^{J} exp\{d_{ij} \log \mu_{ij} \, - \mu_{ij} \},$$

dove $\mu = [\mu_{11}, \mu_{12}, \dots, \mu_{1j}, \dots, \mu_{ij}, \dots]$. (2.18)

Il termine $\frac{1}{d_{ij}!}$ può essere tralasciato in quanto è una costante non dipendente dal parametro.

La relativa log-verosimiglianza si può ottenere applicando la trasformazione logaritmica all'espressione precedente:

$$l(\mu) = \log L(\mu) = \sum_{i=1}^{n}\sum_{j=1}^{J} d_{ij} \log \mu_{ij} \; - \mu_{ij} \tag{2.19}$$

Prendendo in considerazione la relazione (2.11), si può sostituire $\mu_{ij} = \; t_{ij}\lambda_{ij}$:

$$l(\mu) = \sum_{i=1}^{n}\sum_{j=1}^{J} d_{ij} \log \mu_{ij} \; - \mu_{ij} \; = \sum_{i=1}^{n}\sum_{j=1}^{J} d_{ij} \left[\log t_{ij} + log\lambda_{ij}\right] \; - t_{ij}\lambda_{ij} \tag{2.20}$$

Il termine $d_{ij} log t_{ij}$ può essere eliminato in quanto costante, ottenendo così l'espressione della log-verosimiglianza per il modello Poisson applicato nel campo della sopravvivenza:

$$l(\mu) = \sum_{i=1}^{n} l_i\left[\mu_{11}, \mu_{12}, \dots, \mu_{1j}, \dots, \mu_{ij}, \dots\right] =$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{J} l_{ij} = \sum_{i=1}^{n}\sum_{j=1}^{J} d_{ij} \left[\log t_{ij} + log\lambda_{ij}\right] \; - t_{ij}\lambda_{ij} \; \propto$$

$$\propto \sum_{i=1}^{n}\sum_{j=1}^{J} d_{ij} \, log\lambda_{ij} - t_{ij}\lambda_{ij} \tag{2.21}$$

È opportuno definire una leggera modifica alla log-verosimiglianza appena definita, sostituendo l'indice della seconda sommatoria J con $j(i)$. Quest'ultimo termine indica l'intervallo j in cui il soggetto i-esimo cade, ovvero in cui l'individuo sperimenta l'evento o viene censurato.

È possibile dimostrare che la funzione di verosimiglianza per dati censurati a destra generati dalla (2.10) e quella del modello Poisson presentato nella (2.21) coincidono, portando a ottenere le medesime stime dei parametri e procedure inferenziali, giustificando inoltre l'appropriatezza del metodo nel campo della sopravvivenza.

Per dimostrare questa coincidenza, si procede a definire il contributo dell'i-esimo soggetto nella log-verosimiglianza per dati censurati, che si ricava a partire dalla funzione di verosimiglianza calcolata per tutti i soggetti:

$$L = \prod_{i=1}^{n} L_i = \prod_{i=1}^{n} \lambda(t_i)^{d_i} S(t_i). \qquad (2.22)$$

Riportando l'espressione (2.22) in forma logaritmica e tenendo a mente la relazione tra la funzione di sopravvivenza $S(t)$ e la funzione di rischio integrato $H(t)$ vista nella (2.7), si ottiene il contributo del soggetto i-esimo:

$$logL_i = d_i log\lambda_i(t_i) - H_i(t_i), \qquad (2.23)$$

dove $\lambda_i(t)$ e $H_i(t)$ rappresentano la funzione di rischio e la funzione di rischio integrato, rispettivamente, per l'individuo i-esimo al tempo t.

Il primo termine dell'espressione (2.23) può essere riscritto sotto il modello esponenziale a rischi proporzionali costanti a tratti come segue:

$$d_i log\lambda_i(t_i) = d_{ij(i)} log\lambda_{ij(i)}. \quad (2.24)$$

Il termine $\lambda_{ij(i)}$ rappresenta il rischio per l'individuo i-esimo di sperimentare l'evento di interesse nell'intervallo j quando t_i cade nell'intervallo $j(i)$. L'indicatore $d_{ij(i)}$ indica se l'evento, per il soggetto i, si è verificato o meno nell'ultimo intervallo visitato dallo stesso. La funzione di rischio integrato che compare nel secondo termine della (2.23), è l'integrale della funzione di rischio tra 0 e t_i. Tuttavia, questa può essere riscritta come una somma di integrali, uno specifico per ogni intervallo laddove il rischio è costante. Formalmente,

$$H_i(t_i) = \int_0^{t_i} \lambda_i(t)dt = \sum_{j=1}^{j(i)} t_{ij}\lambda_{ij}, \quad (2.25)$$

dove t_{ij} è il tempo che il soggetto i trascorre nell'intervallo j. Il contributo all'integrale per l'intervallo j-esimo sarà il prodotto $t_{ij}\lambda_{ij}$, ovvero il rischio di subire l'evento di interesse per l'ampiezza dell' intervallo stesso. Se l'individuo sperimenta l'evento o è censurato, allora il proprio contributo sarà rappresentato dal rischio moltiplicato per il tempo trascorso dall'inizio dell'intervallo fino alla propria censura o cessazione del rischio, pari a $t_i - \tau_{j-1}$.

Come si può notare nell'espressione (2.25), la funzione di rischio integrato porta ad avere un numero $j(i)$ di termini, ciascuno specifico per ogni j-esimo intervallo e pari al numero di intervalli visitati dal soggetto i-esimo. In precedenza era stato detto che la componente d_{ij} assume valore unitario se l'evento per un determinato soggetto si verifica nell'intervallo j-esimo, assumendo così valore nullo in tutti gli intervalli a questo precedenti. Gli individui per la quale d_{ij} risulterà essere nullo non contribuiranno in nessun modo alla componente della log-verosimiglianza nella (2.24) e quindi si può riscrivere la funzione di log-verosimiglianza complessiva in un modo più semplice e intuitivo, cioè come una somma di $j(i)$ contributi, uno per ogni intervallo. In formule, si ha:

$$l_i = \log L_i = \sum_{j=1}^{j(i)} \{d_{ij} \log \lambda_{ij} - t_{ij}\lambda_{ij}\}. \quad (2.26)$$

Il fatto che il contributo dell'i-esimo individuo alla log-verosimiglianza sia una somma di un certo numero di termini porta a trattare ciascun termine come rappresentativi di un'osservazione indipendente. La funzione di log-verosimiglianza appena presentata è molto simile a quella che si otterrebbe se le d_{ij} avessero una distribuzione Poisson di media $\mu_{ij} = t_{ij}\lambda_{ij}$, eccetto per una costante:

$$l_{ij} = \log L_{ij} = d_{ij} \log \mu_{ij} - \mu_{ij} = d_{ij} \log(t_{ij}\lambda_{ij}) - t_{ij}\lambda_{ij}. \quad (2.27)$$

Questa espressione differisce da quella della log-verosimiglianza (2.26) per il termine $d_{ij} \log t_{ij}$. Tuttavia questo termine può essere tranquillamente tralasciato, poiché è una costante che dipende dai dati e non dai parametri.

I passaggi eseguiti per dimostrare la coincidenza tra le due funzioni di verosimiglianza, possono essere riscritti in una forma più generale, dato che la formulazione precedente riporta il contributo per ciascuna i-esima osservazione. Procedendo con le medesime considerazioni fatte in precedenza, si ottiene:

$$L(\mu) = \prod_{i=1}^{n} L_i = \prod_{i=1}^{n} \lambda(t_i)^{d_i} S(t_i) \quad (2.28)$$

$$l(\mu) = logL(\mu) = \sum_{i=1}^{n} d_i log\lambda_i(t_i) - H_i(t_i) \quad (2.29)$$

I termini $d_i log\lambda_i(t_i)$ e $H_i(t_i)$ possono essere riscritti come nel (2.24) e nel (2.25). Andando a sostituire:

$$l(\mu) = logL(\mu) = \sum_{i=1}^{n}\sum_{j=1}^{j(i)} d_{ij}\log\left(t_{ij}\lambda_{ij}\right) - t_{ij}\lambda_{ij}$$

$$l(\mu) \propto \sum_{i=1}^{n}\sum_{j=1}^{j(i)} d_{ij}\log\lambda_{ij} - t_{ij}\lambda_{ij} \qquad (2.30)$$

Il risultato presentato dimostra che vi è una stretta connessione tra la verosimiglianza di un modello Poisson e quella relativa a osservazioni censurate a destra di un modello esponenziale a rischi proporzionali e costanti a tratti, che sono tipiche di un'analisi della sopravvivenza. Questa particolare relazione è giustificata dal fatto che, anziché avere un singolo indicatore per ogni individuo che ci informa se questo ha sperimentato l'evento di interesse o meno, si dispone di indicatori specifici per ogni intervallo "visitato" da ciascun individuo.

2.5 Covariate ed effetti tempo-dipendenti nel modello di Poisson per dati di sopravvivenza

Nel modello di Poisson specificato in precedenza possono essere introdotte variabili tempo-dipendenti, ovvero un vettore di covariate x_{ij} che rappresentano le caratteristiche di uno specifico individuo i , i cui valori variano in corrispondenza di ciascun intervallo j. Il modello può essere riscritto in maniera semplice come segue:

$$log\mu_{ij} = logt_{ij} + x'_{ij}\,\beta + \sum_{j=1}^{J} I_j\beta_{0j}\,,\ \text{con}\ j = 1, \dots, J. \qquad (2.31)$$

Le variabili esplicative cambiano il proprio valore solamente da un intervallo all'altro e questo, sotto un certo punto di vista, potrebbe sembrare restrittivo. In realtà il risultato consiste in un modello caratterizzato da un'elevata flessibilità, poiché vi è la possibilità di dividere le pseudo osservazioni e di procedere come segue: si considera la covariata di un certo soggetto i che varia in uno specifico intervallo j e si va a dividere l'osservazione stessa in due parti: alla prima viene associato il vecchio valore della covariata, mentre alla seconda quello nuovo; a queste due nuove osservazioni vengono successivamente associate una misura di esposizione t_{ij} e un indicatore binario d_{ij} che suggerisce se l'evento di interesse si sia verificato o meno. Entrambe apparterranno allo stesso intervallo cosicché avranno lo stesso rischio di base $I_j\beta_{0j}$.

Oltre alla variabili tempo-dipendenti possono essere introdotte nel modello anche effetti tempo-dipendenti, denominati come β_j. L'espressione del modello diventa:

$$log\mu_{ij} = logt_{ij} + x'_{ij}\,\beta_j + \sum_{j=1}^{J} I_j\beta_{0j}\,,\text{con}j = 1, \dots, J. \qquad (2.32)$$

β_j rappresenta l'effetto della variabile esplicativa sul rischio nell'intervallo j-esimo.

2.6 Troncamento a sinistra

Negli studi di sopravvivenza si è soliti seguire un gruppo di soggetti per un determinato periodo di tempo, definito come periodo di follow-up, procedendo nel corso dello studio a identificare coloro i quali sperimentano l'evento di interesse. Nel nostro caso si segue un gruppo di pazienti norvegesi affetti da cancro al seno, per un periodo pari a 26 anni. Nel campione osservato, si procede ad arruolare quei soggetti ai quali è stato diagnosticato il tumore in un arco di tempo che intercorre tra il 1965 e il 1974, rilevando alcune caratteristiche di interesse, quali l'età all'insorgenza della neoplasia. Pertanto, si è proceduto a escludere dallo studio tutti gli individui ai quali è stata diagnosticata la stessa tipologia di cancro in anni che precedono l'intervallo di tempo considerato. Questo particolare modo di procedere viene definito come troncamento a sinistra (*left truncation*) e differisce dal meccanismo di censura a sinistra (*left censoring*). Infatti, la prima avviene quando si procede a escludere quegli individui che hanno già sperimentato l'evento prima dell'inizio dello studio o il valore di una certa variabile di interesse non è compresa in un intervallo considerato (in questo caso l'anno della diagnosi del cancro). Il secondo occorre quando si decide di includere ugualmente questi soggetti nello studio censurandoli a sinistra, poiché l'anno in cui gli è stata diagnosticata la malattia è precedente al 1965 e quindi non la si conosce con esattezza. Nel lavoro che verrà presentato non ci saranno casi di censure a sinistra. Da studi condotti tramite simulazioni, si è dimostrato come il fatto di tener conto del troncamento a sinistra può evitare una considerevole distorsione verso il basso delle stime dei parametri di interesse e dei relativi standard errors, anche se i risultati sono instabili quando la frazione di "troncati" è molto elevata (Cain, et al., 2011).

CAPITOLO III

LO STIMATORE NON DISTORTO IN MEDIANA

3.1 Introduzione

Nel terzo capitolo della tesi si affronta un argomento che differisce da quello discusso nelle pagine precedenti, ovvero dall'analisi della sopravvivenza e dei vari modelli presentati che possono essere applicati nel contesto. L'intento è quello di presentare un nuovo stimatore, per il quale vale la proprietà di non distorsione in mediana. La decisione di inserire questa particolare aggiunta, che si discosta in parte da quello che è il tema centrale del lavoro, risiede sia nella volontà di presentare uno strumento innovativo e non di comune trattazione sia nel procedere a un confronto di tipo pratico.

La formalizzazione del capitolo si basa in parte sul lavoro di (Kenne Pagui, Salvan, & Sartori, 2017) e si è deciso di strutturarlo in diversi paragrafi per arrivare ad avere una chiara comprensione di questo particolare stimatore, il quale si ricava fondamentalmente da una modifica della funzione punteggio. In primis verranno richiamate le principale definizioni e concetti della verosimiglianza, soffermandosi in particolare sulle caratteristiche e proprietà dello stimatore di massima verosimiglianza e della funzione punteggio. Successivamente si riprenderanno in mano alcuni concetti che saranno utili nel capire le formulazioni e le quantità che vengono usate per la derivazione della funzione punteggio modificata e quindi del conseguente stimatore. Infine si giungerà al cuore del capitolo, ovvero lo stimatore non distorto in mediana, arrivando a definirlo in tre contesti differenti: per un parametro scalare di interesse in presenza e in assenza di parametri di disturbo e per un vettore di parametri di interesse.

Tutte le definizioni che verranno presentate in questo capitolo fanno riferimento ai testi di (Pace & Salvan, 2001), (Azzalini, 2001) e all'articolo (Kenne Pagui, Salvan, & Sartori, 2017).

3.2 Principali concetti della verosimiglianza

Sia $\mathcal{F}$ un modello statistico parametrico per i dati y con funzione del modello $p_Y(y;\theta)$, con $\theta = (\theta_1, \dots, \theta_p) \in \Theta \subseteq \mathbb{R}^p$. Si consideri $p_Y(y;\theta)$ come funzione solamente del parametro θ, con y fissato al valore osservato. La funzione $L : \Theta \to \mathbb{R}^+$ definita da:

$$L(\theta) = p_Y(y;\theta), \tag{3.1}$$

è detta **funzione di verosimiglianza** (*likelihood function*) di θ basata sui dati y. Verrà utilizzata la scrittura $L(\theta;y)$ quando occorrerà mettere in evidenza la dipendenza di $L(\theta)$ dai dati (Pace & Salvan, 2001, p. 129). Nell'espressione (3.1) è possibile trascurare tutti quei fattori che non dipendono dal parametro θ. Due funzioni di verosimiglianza che differiscono per una costante moltiplicativa si dicono **equivalenti**.

Se si ha a che fare con osservazioni che derivano da variabili casuali indipendenti e identicamente distribuite ($i.i.d.$), con funzione del modello per le distribuzioni marginali $p_{Y_i}(y_i;\theta)$, allora la quantità (3.1) può essere riscritta come segue:

$$L(\theta) = \prod_{i=1}^{n} p_{Y_i}(y_i;\theta), \tag{3.2}$$

ovvero, la funzione di verosimiglianza complessiva è semplicemente il prodotto delle funzioni di verosimiglianza ottenute nei singoli esperimenti.

Spesso si preferisce esprimere le procedure di inferenza basate sulla verosimiglianza $L(\theta)$ attraverso la funzione di **log-verosimiglianza** (*log-likelihood function*),

$$l(\theta) = \log L(\theta). \tag{3.3}$$

La trasformazione monotona crescente logaritmica applicata alla funzione di verosimiglianza offre particolari vantaggi sia a livello computazionale, per l'esecuzione pratica di calcoli, che di interpretazione dei risultati teorici ottenuti. Due funzioni di log-verosimiglianza che differiscono per una costante additiva si dicono equivalenti (Pace & Salvan, 2001, p. 130).

Sotto campionamento casuale semplice, l'espressione (3.3) viene riscritta come:

$$l(\theta) = \sum_{i=1}^{n} \log p_{Y_i}(y_i; \theta). \tag{3.4}$$

Le definizioni che verranno riportate da questo punto in avanti valgono nell'ipotesi in cui si tratti di modelli statistici con **verosimiglianza regolare**. Sotto queste assunzioni, si possono definire le seguenti **quantità di verosimiglianza** (Pace & Salvan, 2001, p. 132-135,139):

- **Funzione punteggio** (*o funzione score*): indicata con $l_*(\theta)$, è il vettore delle derivate parziali prime della funzione di log-verosimiglianza $l(\theta)$, il cui generico elemento è $l_r = \frac{\partial l(\theta)}{\partial \theta_r}$, con $r = 1, \dots, p$;

- **Informazione osservata**: matrice di dimensione $p\,x\,p$ delle derivate parziali seconde di $l(\theta)$ cambiate di segno, in formule:

$$j(\theta) = -l_{**}(\theta) = \left[-\frac{\partial^2 l(\theta)}{\partial \theta_r \partial \theta_s}\right]; \tag{3.5}$$

- **Informazione attesa** (*o informazione di Fisher*): è il valore atteso dell'informazione osservata $j(\theta)$, indicata con:

$$i(\theta) = E_\theta\big(j(\theta)\big) = \left[-E_\theta\left(\frac{\partial^2 l(\theta)}{\partial\theta_r\partial\theta_s}\right)\right]. \qquad (3.6)$$

Per determinare le distribuzioni asintotiche della funzione punteggio e dello stimatore di massima verosimiglianza, si assume che il supporto di Y non dipenda da $\theta \in \Theta \subseteq \mathbb{R}^p$ e che siano soddisfatte le condizioni che permettono di scambiare le operazioni di derivazione e integrazione (Pace & Salvan, 2001, p. 212).

Per n sufficientemente grande, la funzione punteggio $l_*(\theta)$ valutata al vero valore del parametro θ ha distribuzione approssimata:

$$l_*(\theta) \dot{\sim} N\big(0, i(\theta)\big). \qquad (3.7)$$

Un valore $\theta \in \Theta$ che massimizza $L(\theta)$ o in modo equivalente $l(\theta)$, è detto **stima di massima verosimiglianza** (*o s.m.v.*) di θ ed è indicata con $\hat{\theta}$; $\hat{\theta}_n$ rappresenta il relativo **stimatore di massima verosimiglianza**. In un modello caratterizzato da verosimiglianza regolare, $\hat{\theta}$ va cercato tra le soluzioni dell'**equazione di verosimiglianza** $l_*(\theta) = 0$. Nel seguito si assume che tale soluzione sia unica, almeno al divergere della numerosità campionaria n. Lo stimatore di massima verosimiglianza $\hat{\theta}_n$ ha distribuzione approssimata sotto θ:

$$\hat{\theta}_n \dot{\sim} N_p\left(\theta, j\big(\hat{\theta}\big)^{-1}\right). \qquad (3.8)$$

Nell'espressione precedente, la quantità $j\big(\hat{\theta}\big)$ può essere sostituita dalla stima di $i\big(\hat{\theta}\big)$ o da $i(\theta)$. Inoltre lo stimatore di verosimiglianza è asintoticamente non distorto (nei campioni finiti è in generale distorto e con distorsione di ordine $O(n^{-1})$) e asintoticamente efficiente, nel senso che la varianza della sua distribuzione asintotica raggiunge il limite inferiore di Cramér-Rao (Pace & Salvan, 2001, p. 219).

Per indicare le quantità di verosimiglianza valutate in $\hat{\theta}$ verranno usate le seguenti notazioni: $l_*\big(\hat{\theta}\big)$, $j\big(\hat{\theta}\big)$ e $i\big(\hat{\theta}\big)$.

3.3 Proprietà di invarianza ed equivarianza

Le funzioni di verosimiglianza e log-verosimiglianza non dipendono dalla parametrizzazione scelta per il modello statistico $\mathcal{F}$. Sia $\psi = \psi(\theta)$, con $\psi(\cdot)$ funzione biunivoca e regolare definita da $\Theta \subseteq \mathbb{R}^p$ in $\Psi \subseteq \mathbb{R}^p$, una parametrizzazione alternativa del modello e $\theta(\psi)$ la sua funzione inversa. Poiché θ e $\psi(\theta)$ individuano lo stesso elemento di $\mathcal{F}$, si può scrivere:

$$\mathrm{L}^{\Psi}(\psi) = \mathrm{L}^{\Theta}\big(\theta(\psi)\big), \tag{3.9}$$

$$\mathrm{l}^{\Psi}(\psi) = \mathrm{l}^{\Theta}\big(\theta(\psi)\big). \tag{3.10}$$

Le quantità $\mathrm{L}^{\Psi}(\cdot)$ e $\mathrm{l}^{\Psi}(\cdot)$ rappresentano la verosimiglianza e la log-verosimiglianza, rispettivamente, nella parametrizzazione ψ e $\mathrm{L}^{\Theta}(\cdot)$ e $\mathrm{l}^{\Theta}(\cdot)$ le stesse funzioni nella parametrizzazione originaria θ.

Le formulazioni appena viste sono del tutto equivalenti e la scelta tra l'una o l'altra è in parte di convenienza; infatti è possibile scegliere una certa parametrizzazione per rendere più semplice la trattazione matematica che avviene nella fase di inferenza oppure per avere una più chiara interpretazione dei parametri stimati, quando l'obiettivo è quello di descrivere il fenomeno di interesse. La scelta della parametrizzazione non è univoca e quindi ciò che si vuole è che le conclusioni inferenziali siano invarianti rispetto alla parametrizzazione adottata (Azzalini, 2001, p. 63-64).

Quanto appena enunciato, viene definita come **proprietà di invarianza** ed è soddisfatta da tutte le procedure che si basano sul metodo della verosimiglianza. Lo stimatore di massima verosimiglianza, la cui distribuzione è riportata nella (3.8), è caratterizzato dalla **proprietà di equivarianza**, secondo la quale se $\psi(\cdot)$ è una funzione biunivoca regolare dallo spazio Θ a Ψ, allora la stima di massima verosimiglianza di $\psi(\theta)$ è $\psi\big(\hat{\theta}\big)$, con $\hat{\theta}$ stima di massima verosimiglianza di θ relativa a $L(\theta)$. Le due proprietà appena definite non sono da considerare sinonimi e non devono essere confuse l'una con l'altra. Infatti l'equivarianza della s.m.v. rispetto a trasformazioni del parametro, assicura l'invarianza delle conclusioni inferenziali, ma il viceversa non sempre viene garantito.

3.4 Definizione di momenti e cumulanti

Un ingrediente che risulta essere necessario per lo stimatore che verrà presentato nel capitolo, è il cumulante. Per giungere a una sua semplice definizione, si è deciso di partire da quantità che sono ormai ben note nel mondo della statistica: i momenti.

Lo studio della distribuzione attraverso i suoi momenti è vantaggioso perché permette di ottenere direttamente informazioni, quanto più dettagliate, sulla forma della distribuzione stessa. Tuttavia, in alcuni contesti vi sono delle quantità connesse ai momenti di una distribuzione che risultano essere più convenienti da usare, ovvero i cumulanti. Essi descrivono l'informazione guadagnata all'ordine $r-esimo$, che non era presente agli ordini precedenti (Pace & Salvan, 1997, p. 72-80)

Sia Y una variabile casuale unidimensionale avente funzione di densità $p_Y(y)$ e siano $\mu_r = E(Y^r)$ i suoi momenti, con r = 1,2,… . La funzione generatrice dei momenti di Y (*m.g.f.*) viene definita come:

$$M_Y(t) = E(e^{tY}), \quad con\ t \in \mathbb{R}. \tag{3.11}$$

Se $M_Y(t)$ esiste finita per $|t| < t_0$, con $t_0 > 0$, allora Y ha momenti finiti per ogni suo ordine e tale risultato può essere dimostrato attraverso l'espansione in serie di potenze di $M_Y(t)$:

$$M_Y(t) = 1 + \mu_1 t + \mu_2 \frac{t^2}{2!} + \mu_3 \frac{t^3}{3!} + \mu_4 \frac{t^4}{4!} + \cdots \tag{3.12}$$

Dallo sviluppo presentato nella (3.12) è possibile ricavare l'espressione del generico momento di ordine r, tramite la relazione $\mu_r = \frac{d^r}{dt^r} M_Y(t)|_{t=0}$.

Se $M_Y(t)$ è finita per $|t| < t_0$ con $t_0 > 0$, si arriva a definire la funzione generatrice dei cumulanti (*c.g.f.*) di Y:

$$K_Y(t) = \log M_Y(t). \tag{3.13}$$

In modo del tutto analogo a prima, la funzione $K_Y(t)$ può essere espressa come uno sviluppo in serie di potenze, del tipo:

$$K_Y(t) = k_1 t + k_2 \frac{t^2}{2!} + k_3 \frac{t^3}{3!} + k_4 \frac{t^4}{4!} + \cdots \tag{3.14}$$

Dove il coefficiente k_r del generico elemento r dello sviluppo, viene definito come **cumulante di ordine r** di Y e si calcola come: $k_r = \frac{d^r}{dt^r} K_Y(t)|_{t=0}$.

I cumulanti possono essere riscritti in funzione dei momenti o dei momenti centrali:

$$\begin{aligned} k_1 &= \mu_1, \\ k_2 &= \mu_2 - \mu_1^2 = \bar{\mu}_2 = Var(Y), \\ k_3 &= \mu_3 - 3\mu_1\mu_2 + 2\mu_1^3 = \bar{\mu}_3, \\ k_4 &= \mu_4 - 3\mu_2^2 - 4\mu_1\mu_3 + 12\mu_1^2\mu_2 - 6\mu_1^4 = \bar{\mu}_4 - 3\bar{\mu}_2^2. \end{aligned} \tag{3.15}$$

I primi quattro cumulanti rappresentano le misure descrittive di posizione, variabilità, asimmetria e curtosi, rispettivamente. Cumulanti aventi ordine superiore forniscono ulteriori informazioni se confrontati con i rispettivi della distribuzione Normale.

Infine, se $S_n = \sum_{i=1}^{n} Y_i$, con Y_i realizzazioni indipendenti e identicamente distribuite di Y, allora è possibile ottenere le funzioni generatrici dei momenti e dei cumulanti di S_n tramite le relazioni:

$$\mathrm{M_{S_n}}(\mathrm{t}) = (M_Y(t))^n \quad \text{e} \qquad K_{\mathrm{Y}}(t) = nK_Y(t) \tag{3.16}$$

3.5 Lo stimatore non distorto in mediana

La non distorsione rappresenta una proprietà desiderabile per uno stimatore e fornisce un giudizio circa la bontà dello stesso. In particolare, la maggior parte degli stimatori risultano essere non distorti in media. Nei contesti regolari di stima, lo stimatore di massima verosimiglianza $\hat{\theta}_n$ e la funzione punteggio $l_*(\theta)$ sono caratterizzati da una distribuzione asintotica e simmetrica, centrata sul vero valore del parametro θ e sullo 0, rispettivamente (si veda (3.7) e (3.8)).

Tuttavia le proprietà asintotiche possono riflettere in maniera erronea il comportamento delle distribuzioni per campioni finiti. Per trattare il problema della distorsione sono state proposte differenti soluzioni in letteratura, molte delle quali cercano di aggiustare in maniera approssimata la distorsione, sia dello stimatore di massima verosimiglianza che della funzione punteggio quando sono presenti parametri di disturbo. Nel primo caso si faccia riferimento a (Kosmidis, 2014) e nel secondo a (McCullagh & Tibshirani, 1990) e (Stern, 1997). In assenza di parametri di disturbo, la funzione punteggio $l_*(\theta)$ è esattamente non distorta in media, di conseguenza non è necessario apportare alcun aggiustamento. Nel contesto in cui si decidesse di cambiare parametrizzazione, la non distorsione della funzione punteggio resterebbe comunque garantita, in quanto la soluzione dell'equazione di verosimiglianza, $l_*(\theta) = 0$, ovvero la stima di massima verosimiglianza $\hat{\theta}$, è equivariante rispetto a trasformazioni del parametro, assicurando quindi l'invarianza delle conclusioni inferenziali (si veda paragrafo (3.2)). (Firth, 1993), a differenza degli autori precedenti, ha proposto una modifica della funzione punteggio per ridurre la distorsione dello stimatore di massima verosimiglianza e garantire la finitezza delle stime in quei casi in cui $\hat{\theta}$ risulta essere infinito.

Nell'ambito della tesi, questi metodi presentati brevemente non saranno trattati, poiché il vero scopo di quest'ultimo capitolo risiede nel giungere a una definizione di un nuovo stimatore, garantendone la non distorsione in mediana piuttosto che quella in media. La metodologia utilizzata in questo capitolo, deriva dal lavoro di (Kenne Pagui, Salvan, & Sartori, 2017).

Considerando un parametro scalare di interesse θ, verrà inizialmente presentata una modifica dell'equazione di verosimiglianza $l_*(\theta) = 0$, la cui soluzione soddisfa la proprietà di equivarianza sotto riparametrizzazioni monotone. Così come in Firth, questa nuova modifica introdotta non dipende dalle stime di massima verosimiglianza e quindi può fornire stime finite anche quando $\hat{\theta}$ non lo è. La modifica proposta è ottenuta considerando la mediana come indice di centralità per la funzione punteggio $l_*(\theta)$, invece della media, arrivando a definire una nuova funzione di stima sottraendo a $l_*(\theta)$ l'espressione della sua mediana approssimata.

Verrà provato che l'equazione modificata è caratterizzata da un'unica soluzione e il fatto di centrare la funzione punteggio sulla mediana implica la centratura in mediana del corrispondente stimatore. Perciò lo stimatore che ne risulterà sarà approssimatamene non distorto in mediana al terzo ordine, ovvero lo stimatore avrà eguale probabilità di sovrastimare o sottostimare il vero valore del parametro. In alcuni casi è possibile ottenere stime esatte in mediana (Hirji, Tsiatis, & Mehta, 1989). Le approssimazioni disponibili per le stime non distorte in mediana sono basate su risultati asintotici di verosimiglianza di ordine superiore e sul rapporto di verosimiglianza, proposti da (Barndorff-Nielsen, 1986), (Pace & Salvan, 1999), (Guimmole & Ventura, 2002), (Biehler, Holling, & Doeber, 2015). Tutte queste approssimazioni forniscono stime finite della stima di massima verosimiglianza.

Questo metodo può essere eventualmente esteso a un vettore di parametri di interesse, risolvendo simultaneamente le equazioni di verosimiglianza non distorte in mediana per ciascun parametro. Ciò porterà alla non distorsione in mediana al terzo ordine per ogni componente e all'equivarianza rispetto a riparametrizzazioni.

Da questo punto in avanti la funzione punteggio verrà indicata semplicemente come $U(\theta)$, per non creare confusione con la notazione di riferimento.

3.5.1 La funzione di verosimiglianza modificata in mediana per un parametro scalare di interesse, in assenza di parametri di disturbo

Si consideri un modello statistico con verosimiglianza regolare per i dati y, avente funzione di densità $p_Y(y;\theta)$ e con parametro scalare di interesse $\theta \in \Theta \subseteq \mathbb{R}$. Sia $l(\theta)$ la corrispondente funzione di log-verosimiglianza e $U(\theta)$ la funzione punteggio. $\hat{\theta}$ rappresenta la stima di massima verosimiglianza, ossia la soluzione dell'equazione di verosimiglianza $U(\theta) = 0$. Si assuma infine che $i(\theta)$ sia l'informazione osservata e che i cumulanti fino al terzo ordine di $l(\theta)$ risultino essere finiti e dell'ordine di $O(n)$, con n indicante la dimensione campionaria.

Per spiegare la modifica dell'equazione di verosimiglianza e quindi la definizione dello stimatore, si parte da un'espansione asintotica che approssima la mediana della funzione punteggio $U(\theta)$, per il parametro θ nel caso continuo. Tale espressione è nota come espansione di Cornish-Fisher (Pace & Salvan, 1997, p. 394):

$$y_\alpha = z_\alpha + \frac{\rho_3}{6\sqrt{n}}(z_\alpha^2 - 1) + \frac{\rho_4}{24n}(z_\alpha^3 - 3z_\alpha) - \frac{\rho_\alpha^3}{36n}(2z_\alpha^3 - 5z_\alpha) + O\left(n^{-\frac{3}{2}}\right), \quad (3.17)$$

dove y_α è il quantile $\alpha - esimo$ di ${S_n}^*$, con ${S_n}^* = (S_n - n\mu)/\sqrt{n\sigma^2}$. Il termine z_α rappresenta i quantile della Normale Standard, mentre ρ_r sono i cumulanti standardizzati, definiti da: $\rho_r = \frac{k_r}{k_r^{r/2}}$.

L'espansione di Cornish-Fisher può essere utilizzata per l'approssimazione dei quantili di una distribuzione opportunamente scelta, tuttavia, come è stato detto in precedenza, questo metodo sarà applicato per approssimare la mediana della funzione punteggio o della funzione punteggio profilo, fissando opportunamente il parametro α pari a 0.5. Sfruttando la (3.17), si può ottenere infatti l'approssimazione cercata. In formule:

$$M_\theta\{U(\theta)\} = \frac{-v_{\theta,\theta,\theta}}{\{6\, i(\theta)\}} + O(n^{-1}), \tag{3.18}$$

con $v_{\theta,\theta,\theta} = v_{\theta,\theta,\theta}(\theta) = E_\theta\{U(\theta)^3\}$.

Uguagliando la funzione Punteggio $U(\theta)$ al termine principale della sua mediana, specificata nella (3.18), si giunge alla modifica proposta inizialmente. In formule:

$$\tilde{U}(\theta) = U(\theta) + \frac{v_{\theta,\theta,\theta}}{\{6\, i(\theta)\}}, \tag{3.19}$$

dove il termine di modifica $\frac{v_{\theta,\theta,\theta}}{\{6\, i(\theta)\}}$ è dell'ordine di $O(1)$.

Sia $\tilde{\theta}_n$ lo stimatore di massima verosimiglianza ottenuto come soluzione dell'equazione di verosimiglianza modificata $\tilde{U}(\theta) = 0$. Per costruzione vale che $M_\theta\{\tilde{U}(\theta)\} = O(n^{-1})$ e, per monotonicità di $\tilde{U}(\theta)$, $P_\theta\{\tilde{U}(\theta) \leq 0\} = \frac{1}{2} + O\left(n^{-\frac{3}{2}}\right)$, implica che se la stima di massima verosimiglianza ottenuta dall'equazione modificata è unica, allora $P_\theta\{\tilde{\theta}_n \leq \theta\} = \frac{1}{2} + O\left(n^{-\frac{3}{2}}\right)$, da cui segue che lo stimatore $\tilde{\theta}_n$ è non distorto in mediana al terzo ordine.

Come $\hat{\theta}_n$, anche $\tilde{\theta}_n$ possiede una distribuzione asintotica del tipo:

$$\tilde{\theta}_n \dot{\sim} N(\theta, i(\theta)^{-1}). \tag{3.20}$$

Poiché lo stimatore che si ottiene dall'equazione di verosimiglianza modificata ha una distribuzione normale asintotica, è quindi possibile procedere alla costruzione di intervalli di confidenza alla Wald, in senso approssimato, che differiscono solo per la posizione rispetto a quelli per $\hat{\theta}_n$, cioè non risulteranno più essere centrati in media, ma in mediana. In alternativa possono essere costruiti intervalli di

confidenza basati sulla funzione punteggio modificata $\tilde{U}(\theta)$, usando la distribuzione asintotica $\tilde{U}(\theta) \dot{\sim} N(0, i(\theta))$.

La nuova proposta ha come principale vantaggio l'essere invariante rispetto a riparametrizzazioni monotone del parametro di interesse, proprietà che risulta essere ereditata dall'invarianza delle statistiche d'ordine. Infatti se ipotizziamo che $\psi(\theta)$, sia una riparametrizzazione monotona con inversa $\theta(\psi)$ e le quantità viste nella (3.19) vengono intese nella nuova parametrizzazione come $v^{\Theta}_{\psi,\psi,\psi} = v_{\theta,\theta,\theta}\{\theta(\psi)\}\{\theta'(\psi)\}^3$ e $i^{\Theta}(\psi) = i\{\theta(\psi)\}\{\theta'(\psi)\}^2$, dove $\theta'(\psi) = \partial\theta(\psi)/\partial\psi$, allora la funzione punteggio modificata $\tilde{U}(\theta)$ si comporta in modo tensoriale di ordine 1 analogamente a $U(\theta)$, cioè la funzione punteggio modificata nella riparametrizzazione monotona ψ è $\tilde{U}\{\theta(\psi)\}\theta'(\psi)$ e di conseguenza, $\tilde{\theta}_n$ è equivariante a trasformazioni monotone del parametro come $\hat{\theta}_n$, e $\tilde{\psi} = \psi(\tilde{\theta}_n)$ è uno stimatore non distorto in mediana fino al terzo ordine.

In un contesto in cui Y appartiene alla famiglia di dispersione esponenziale con parametro canonico θ e funzione di densità $f(y|\,\theta) = h(y)\, exp\{\theta t(y) - K(\theta)\}$, la funzione punteggio modificata si presenta nella forma:

$$\tilde{U}(\theta) = U(\theta) + \frac{K_{\theta\theta\theta}}{6K_{\theta\theta}}, \tag{3.21}$$

dove $K_{\theta\theta\theta} = \partial^3 K(\theta)/\partial\theta^3$ e $K_{\theta\theta} = \frac{\partial^2 K(\theta)}{\partial\theta^2} = i(\theta)$. In questa particolare parametrizzazione, la funzione $\tilde{U}(\theta)$ può essere vista come derivata della log-verosimiglianza penalizzata $\tilde{l}(\theta) = l(\theta) + \{\log i(\theta)\}/6$. Questa espressione dimostra come l'effetto della modifica in mediana, nell'ambito delle famiglie esponenziali, sia quello di penalizzare la verosimiglianza di un fattore pari a $i(\theta)^{1/6}$.

3.5.2 La funzione di verosimiglianza modificata in mediana per un parametro scalare di interesse, in presenza di parametri di disturbo

Nel precedente paragrafo è stato definito il nuovo stimatore non distorto in mediana attraverso una modifica della funzione punteggio, nell'ipotesi in cui si avesse a che fare con un unico parametro scalare di interesse e in assenza di parametri di disturbo. Tuttavia, nella maggior parte dei casi pratici, si lavora con una moltitudine di parametri, alcuni dei quali possono essere considerati dei veri e propri elementi di disturbo per la stima del parametro di interesse. Viene di seguito proposta un'ulteriore modifica della funzione punteggio, che fa fronte a questa seconda problematica.

Sia $\theta = (\theta_1, \dots, \theta_p)$ un vettore di parametri e $U_r = \frac{\partial l(\theta)}{\partial \theta^r}$, con $r = 1, \dots, p$ il generico elemento del vettore della funzione punteggio $U(\theta)$. Sia i_{rs} l'elemento in posizione (r, s) della matrice di informazione attesa $i(\theta)$ e i^{rs} il generico elemento in posizione (r, s) della sua inversa, con $r, s = 1, \dots, p$. Siano inoltre U_{rs} e U_{rst} le derivate parziali di ordine superiore di $l(\theta)$ rispetto agli elementi di θ di indici r, s, t. Si definiscano infine i valori attesi delle derivate della log verosimiglianza con le seguenti notazioni: $v_{rs} = E_\theta(U_{rs}) = -i_{rs}$, $v_{rst} = E_\theta(U_{rst})$, $v_{r,st} = E_\theta(U_r U_{st})$ e $v_{r,s,t} = E_\theta(U_r U_s U_t)$.

Si supponga che il parametro venga partizionato come $\theta = (\psi, \lambda)$, con ψ parametro scalare di interesse. Nel caso in cui si desideri porre l'attenzione sul parametro ψ, di diretto interesse per il ricercatore, e risulti essere possibile eliminare il parametro di disturbo λ, allora si può applicare la metodologia vista al paragrafo precedente (3.5.1) alla funzione punteggio marginale o condizionata di ψ. Nelle maggior parte delle situazioni o quando non è disponibile una soluzione esatta, viene proposta una modifica della funzione punteggio profilo. Sia $l_P(\psi) = l(\psi, \hat{\lambda}_\psi)$ la log verosimiglianza profilo per il parametro ψ, dove $\hat{\lambda}_\psi$ è la stima di massima verosimiglianza di λ per un dato valore di ψ. La funzione punteggio profilo è rappresentata da $U_P(\psi) = \frac{\partial l_P(\psi)}{\partial \psi}$. Per completare la notazione necessaria

a illustrare la modifica proposta, si indichi con gli indici $a, b, c, \ldots$ le componenti di λ, cosicché gli elementi di $U(\theta)$ siano:

- $U_\psi = U_\psi(\psi, \lambda) = \frac{\partial l(\psi,\lambda)}{\partial \psi}$
- $U_a = U_a(\psi, \lambda) = \frac{\partial l(\psi,\lambda)}{\partial \lambda_a}$, con $a = 1, \ldots, p-1$.

Come è risaputo, $U_P(\psi) = U_\psi\left(\psi, \hat{\lambda}_\psi\right)$ e le espressioni approssimate per i primi tre cumulanti di $U_P(\psi)$, tratte da (McCullagh & Tibshirani, 1990) e (Barndorff-Nielsen & Cox, 1989), sono:

$$k_{1\psi} = -\frac{1}{2} v^{ab}\left\{\left(v_{\psi,ab} - \gamma_{\psi c} v_{c,ab}\right) + \left(v_{\psi,a,b} - \gamma_{\psi c} v_{a,b,c}\right)\right\}$$

$$k_{2\psi} = v_{\psi,\psi} - \gamma_{\psi a} v_{\psi,a}$$

$$k_{3\psi} = v_{\psi,\psi,\psi} - 3\gamma_{\psi a} v_{\psi,\psi,a} + 3\gamma_{\psi a}\gamma_{\psi b} v_{\psi,a,b} - \gamma_{\psi a}\gamma_{\psi b}\gamma_{\psi c} v_{a,b,c} \,. \qquad (3.22)$$

Usando l'espansione di Cornish-Fisher vista nel paragrafo precedente, in un modello continuo la mediana della funzione punteggio profilo standardizzata $\left(U_p(\psi) - k_{1\psi}\right)/\sqrt{k_{2\psi}}$ risulta essere pari a $-\frac{k_{3\psi}}{\left\{6k_{2\psi}^{\frac{3}{2}}\right\}} + O\left(n^{-\frac{3}{2}}\right)$. Si ottiene quindi la funzione punteggio profilo modificata in mediana:

$$\widetilde{U}_P(\psi) = U_p(\psi) - k_{1\psi} + \frac{k_{3\psi}}{\{6k_{2\psi}\}}, \qquad (3.23)$$

che ha mediana nulla con errore dell'ordine di $O(n^{-1})$. Allo stesso modo del caso precedente, si mostra che $P_\theta\left\{\widetilde{U}_P(\psi) \leq 0\right\} = \frac{1}{2} + O\left(n^{-\frac{3}{2}}\right)$. Lo stimatore di massima verosimiglianza che si ottiene come soluzione di $\widetilde{U}_P(\psi) = 0$, con $\hat{\lambda}_\psi$ al

posto di λ, è rappresentato da $\tilde{\psi}_p$. Se la stima di massima verosimiglianza è unica, allora lo stimatore $\tilde{\psi}_p$ è non distorto in mediana fino al terzo ordine. Sebbene tale stimatore sia stato definito nel caso continuo, la non distorsione viene garantita anche se vengono presi in considerazione modelli statistici discreti, la cui dimostrazione non viene trattata direttamente nella tesi (Kenne Pagui, Salvan, & Sartori, 2017).

La distribuzione asintotica dello stimatore $\tilde{\psi}_p$ è:

$$\tilde{\psi}_p \sim N\left(\psi, {K_{2\psi}}^{-1}\right). \tag{3.24}$$

Tale distribuzione può essere utilizzata per costruire intervalli di confidenza alla Wald. In alternativa, come nel caso di un unico parametro scalare di interesse, si può utilizzare la distribuzione asintotica della verosimiglianza profilo modificata $\widetilde{U}_P(\psi)$ per gli intervalli di confidenza: $\widetilde{U}_P(\psi) \sim N\left(0, K_{2\psi}\right)$.

Lo stimatore di massima verosimiglianza $\tilde{\psi}_p$ è equivariante rispetto a trasformazioni monotone del parametro di interesse ψ. In dettaglio, se si definisce $\omega = (\varphi, \chi)$ una parametrizzazione monotona con $\varphi = \varphi(\psi)$, $\chi = (\psi, \chi)$ e $\varphi = \varphi(\psi)$ funzione uno-a-uno di ψ con inversa $\psi(\varphi)$, allora la funzione punteggio modificata per φ nella nuova parametrizzazione risulta essere pari a:

$$\widetilde{U}_P\left(\psi(\varphi)\right) \psi'(\varphi), \tag{3.25}$$

dove $\tilde{\varphi}_p = \varphi\left(\tilde{\psi}_p\right)$ sarà il conseguente stimatore di massima verosimiglianza.

Si specifica infine la modifica della funzione punteggio nel contesto in cui si abbia a che fare con un modello statistico appartenente alla famiglia esponenziale di ordine p con parametro canonico (ψ, λ). La funzione di densità è rappresentata da $f(y|\,\psi, \lambda) = h(y)\, exp\{\psi t(y) + \lambda^T s(y) - K(\psi, \lambda)\}$, i cui cumulanti sono ricavati semplicemente dalle derivate di $K(\psi, \lambda)$. La modifica proposta risulta essere $U_p(\psi) - k_{1\psi}$, che consiste in un'approssimazione dell'ordine di $O(n^{-1})$. In precedenza, nel caso continuo lo stimatore di massima verosimiglianza, derivato al (3.23), è un'approssimazione dello stimatore ottimo condizionato non distorto in mediana (Lehmann & Romano, 2005), soluzione rispetto a ψ di

$P_\psi(T \leq t|S = s) = 1/2$. L'approssimazione che si cerca è ottenuta sostituendo $P_\psi(T \leq t|S = s) = 1/2$ con l'approssimazione di Edgeworth (Pace & Salvan, 1997, p. 386) fino ai termini di ordine $O(n^{-1})$.

3.5.3 La funzione di verosimiglianza modificata in mediana per un vettore di parametri di interesse

Nel caso in cui si disponga di un vettore di parametri interesse θ, non è possibile applicare una diretta estensione della (3.19) a causa della mancanza di una definizione esatta di mediana in più dimensioni. Nonostante emergano dalla letteratura alcune possibili definizioni a riguardo (Oja, 2013), nessuna di esse sembra essere adatta per sviluppare una modifica in mediana del vettore della funzione punteggio, o per poca praticabilità o per scarse proprietà campionarie dimostrate in studi simulativi. Si decide di conseguenza di seguire un particolare tipo di approccio, il quale prevede l'impostazione di un sistema di p equazioni, una per ciascun parametro di interesse θ_r, con $r = 1, \dots, p$. In ciascuna delle equazioni, si considera l'r-esimo elemento come unico parametro di interesse e i restanti $p - 1$ parametri come se fossero degli elementi di disturbo, quindi il sistema sarà composto da p funzioni dello stesso tipo di quelle definite al (3.23). Sotto queste assunzioni, è possibile ottenere la seguente funzione punteggio modificata in mediana, per un vettore di parametri di interesse:

$$\breve{U}_r(\theta) = U_r(\theta) - \gamma_{ra} U_a(\theta) + M_r, \quad con\, r = 1, \dots, p \quad (3.26)$$

dove $M_r = -k_{1r} + \frac{k_{3r}}{6k_{2r}}$ e $k_{jr}, j = 1,2,3$ rappresentano i cumulanti di ordine j, che sono gli stessi di quelli visti al (3.22), con $\psi = \theta_r$. Si sottolinea inoltre che i pedici $a, b, c, \dots$, presenti nelle espressioni (3.22) e (3.26), assumono valori in $\{1, \dots, p\} \setminus \{r\}$ e vengono sommati se ripetuti. La stima congiunta dei parametri $\tilde{\theta}$ è ottenuta risolvendo l'equazione $\breve{U}(\theta) = 0$. Concludendo, per ogni $r = 1, \dots, p$, $\breve{U}_r$ si comporta in modo tensoriale sotto riparametrizzazioni dei rispettivi parametri di interesse θ_r e, di conseguenza, $\tilde{\theta}$ è equivariante sotto

riparametrizzazioni congiunte che trasformano ciascuna componente di θ separatamente.

Si indichi con $\bar{U}(\theta)$ l'*efficient score* (Pace & Salvan, 1997), i cui elementi sono: $\bar{U}_r = U_r - \gamma_{ra}U_a$, con $r = 1, \dots, p$. È possibile riscrivere $\bar{U}(\theta)$ come prodotto di una matrice non singolare e non stocastica $A(\theta)$ di ordine p e della funzione punteggio $U(\theta)$, ovvero: $\bar{U}(\theta) = A(\theta)U(\theta)$. Si definisca l'Hessiana come $H(\theta) = E_\theta\left\{-\frac{\partial \bar{U}(\theta)}{\partial \theta^T}\right\} = \{diag(i(\theta)^{-1})\}^{-1}$. Inoltre sia $H(\theta) = A(\theta)i(\theta)$, cosicché si possa ricavare $A(\theta) = H(\theta)i(\theta)^{-1}$ (Kenne Pagui, Salvan, & Sartori, 2017).

Quindi, risolvere l'equazione di verosimiglianza $\tilde{U}(\theta) = 0$ risulta essere equivalente a risolvere:

$$U(\theta) + i(\theta)M_1(\theta) = 0, \tag{3.27}$$

le cui componenti di $M_1(\theta)$ sono $M_{1r}(\theta) = M_r/k_{2r}$. Tuttavia non vi sono garanzie che l'espressione (3.27) abbia una soluzione. Comunque, $i(\theta)M_1(\theta)$ è dell'ordine di $O(1)$ e quindi, in maniera asintotica, l'esistenza della stima di massima verosimiglianza dell'equazione di verosimiglianza modificata $\tilde{\theta}$ è garantita laddove $\hat{\theta}$ esiste. Inoltre, $\tilde{\theta} - \hat{\theta} = O_p(n^{-1})$ e la distribuzione asintotica dello stimatore $\tilde{\theta}_n$ è la stessa di quella dello stimatore $\hat{\theta}_n$.

Sia $\tilde{\theta}_r$ la componente di ordine r di $\tilde{\theta}$ e sia $\tilde{\theta}_{rp}$ la soluzione di $\tilde{U}_p(\theta_r) = 0$, con $\tilde{U}_p(\cdot)$ data dalla (3.23). In un modello regolare, si dimostra che:

$$\tilde{\theta}_r - \tilde{\theta}_{rp} = O_p\left(n^{-\frac{3}{2}}\right), \tag{3.28}$$

con $r = 1, \dots, p$. Ne risulta che la non distorsione in mediana per ogni componente di $\tilde{\theta}_r$ al terzo ordine è implicata dalla proprietà analoga di $\tilde{\theta}_{rp}$.

CAPITOLO IV

APPLICAZIONE DEL MODELLO DI POISSON PER DATI DI SOPRAVVIVENZA

4.1 Il cancro al seno

La World Health Organization (*WHO*) definisce il cancro al seno come la neoplasia che affligge più frequentemente le donne, colpendone oltre un milione e mezzo ogni anno, causando inoltre il maggior numero di morti cancro-correlate. Solamente nel 2015, sono circa 570.000 le donne decedute a causa di questa malattia, le quali rappresentano circa il 15% di tutti i decessi provocati dal cancro. Ad essere più colpite, in genere, sono le donne che appartengono alle regioni più sviluppate del mondo, anche se si osserva un aumento delle incidenze a livello globale. A dispetto di come ci si potrebbe aspettare, il cancro al seno può colpire anche gli uomini, seppur in minor misura rispetto alle donne.

La diagnosi precoce e lo screening rappresentano due strategie ritenute fondamentali per garantire una maggiore sopravvivenza del cancro al seno. Le strategie di diagnosi precoce si preoccupano di garantire un accesso tempestivo al trattamento della neoplasia, riducendo gli ostacoli all'assistenza e migliorando l'accesso ai servizi di diagnosi efficaci. Lo screening, invece, consiste nell'eseguire dei test ai soggetti che sono maggiormente a rischio di sviluppare il cancro al fine di identificarlo prima che si manifestino i sintomi. Vi sono diversi metodi che sono stati valutati come strumenti di screening per il tumore al seno, tra i quali la mammografia, l'esame clinico della mammella (*Clinical Breast Exam o CBE*) e l'autopalpazione. Al giorno d'oggi, la mammografia rappresenta lo strumento più efficace per la prevenzione del tumore al seno, in cui gli individui a rischio di sviluppare la neoplasia vengono chiamati da un'organizzazione centrale per un controllo specifico. Questo metodo permette di ridurre la mortalità causata dal cancro al seno, consentendo inoltre alla sua rimozione per via chirurgica nel caso in cui si riesca a individuare il tumore in fase precoce.

Lo screening richiede investimenti sostanziosi e comporta costi significativi sia dal punto di vista finanziario che da quello del personale. Per questo motivo, la decisione di procedere allo screening deve essere perseguita solo se possono essere garantite determinate condizioni: se a una diagnosi efficace segue un trattamento tempestivo per un intero gruppo di target, se la sua efficacia è stata dimostrata nella regione specifica e se sono disponibili risorse per sostenere il programma e mantenerne la qualità (World Health Organization, 2018).

In questo lavoro verranno presi in considerazione i soggetti, residenti in Norvegia, affetti da neoplasia mammaria. Anche in questo paese, il cancro al seno è la neoplasia più comune tra le donne e, sulla base delle normali aspettative di vita, una donna su dieci svilupperà questa malattia. Nei paragrafi successivi si procederà alla descrizione del dataset e alle prime analisi esplorative.

4.2 Presentazione del dataset

Le analisi che verranno presentate in questo capitolo si basano su dati provenienti dal Registro Tumori Norvegese (Cancer Registry of Norway, 2018). Il Registro fa parte dell'Autorità Sanitaria Regionale della Norvegia sud-orientale (*South-Eastern Norway regional Health Authority*) ed è organizzato come ente indipendente sotto l'Ospedale di Oslo. Consiste in uno dei più antichi registri nazionali di cancro al mondo e il suo utilizzo è molto diffuso sia in ambito nazionale che internazionale. Le informazioni relative al cancro provengono da differenti fonti indipendenti, in particolare dagli ospedali dove i medici sono obbligati per legge a notificare qualsiasi nuovo caso di cancro, anche se la diagnosi non viene accertata. Questo consente al registro un elevato grado di completezza. Una delle principali peculiarità del Registro risiede nei suoi programmi nazionali di screening per la prevenzione del cancro al seno e alla cervice.

Il dataset riporta una serie di informazioni riguardanti la mortalità di soggetti per i quali è stato diagnosticato un tumore al seno in un periodo di tempo che intercorre tra il 1965 e il 1974. I dati raccolti sono relativi a 9041 individui affetti da cancro al seno, per i quali sono state rilevate 14 variabili esplicative.

La variabile risposta considerata è **Status**, dicotomica, la quale indica se il soggetto è deceduto in seguito alla neoplasia o è sopravvissuto/emigrato, attraverso i valori 1 e 0, rispettivamente. Relativamente alle caratteristiche socio-demografiche dell'individuo sono state prese in considerazione le variabili: **Gender** (variabile categoriale che indica se il soggetto appartiene al genere femminile o maschile, tramite le modalità "Female" e "Male", rispettivamente), **Age** (variabile numerica che riporta l'età del paziente) e **age** (un'ulteriore variabile che si riferisce all'età, intesa come differenza tra l'età individuale e la media dell'età di tutti i soggetti nello studio).

La maggior parte delle altre variabili presenti nel dataset fanno riferimento alle caratteristiche tumorali e al periodo di permanenza dell'individuo nello studio: **Surgery1** (variabile categoriale la cui prima modalità, chiamata "No Surgery", suggerisce che non si è proceduto all'eradicazione del tumore per via chirurgica, mentre le restanti tre modalità, rappresentano differenti tipologie di interventi: "Surgery for organs", "Tumor exploration or other interventions" e "Tumor surgery"), **Surgery** (variabile molto simile alla precedente e per questo motivo si è deciso di escluderla dal dataset), **SIDE** (indica il lato del seno in cui è localizzato il tumore), **Metastasis** (variabile categoriale inerente alla presenza di metastasi a distanza, nella regione tumorale o localizzate nella sede del tumore primitivo, rappresentate dalle modalità: "Distant", "Regional" e "Localized" rispettivamente), **Localization** (indica la parte del seno in cui è localizzato il tumore), **Stadium** (rappresenta lo stadio del tumore tramite le modalità "I", "II", "III" e "IV"), **time** (variabile continua che rappresenta il periodo in cui il soggetto viene tenuto sotto osservazione, espressa in giorni), **YEAR_DIAGNOSIS** e **YEAR_STATUS** (variabili che indicano l'anno in cui al soggetto viene diagnosticato il tumore al seno e l'anno della relativa uscita dallo studio per morte/migrazione/fine dello studio, rispettivamente). L'ultima variabile del dataset è **ID**, la quale rappresenta il codice identificativo dell'individuo.

Prima di procedere con le analisi, vengono eliminate le variabili **Surgery** e **SIDE**, poiché questa presenta un'elevata percentuale di valori mancanti (96%).

4.3 Analisi preliminari

Nelle analisi preliminari del fenomeno si è deciso di procedere in due fasi: inizialmente sono state svolte alcune analisi univariate, al fine di ottenere una descrizione generale del campione oggetto di studio; in una fase successiva, è stata approfondita la comprensione delle relazioni tra le variabili, in particolare si è osservato come varia la sopravvivenza dei soggetti affetti da cancro al seno in funzione delle loro caratteristiche socio-demografiche e tumorali.

Prendendo in considerazione la variabile risposta **Status**, si osserva come circa due terzi dei soggetti nel campione siano morti a causa della neoplasia (68%), mentre i restanti sono sopravvissuti o emigrati. La popolazione non è distribuita equamente per genere, in quanto la quasi totalità del campione analizzato è costituito da individui di sesso femminile (8995 femmine contro 46 maschi). Discriminando le morti per genere (*Grafico 4.1*), si osserva come la percentuale dei decessi relativi ai maschi sia più elevata rispetto alla controparte femminile, anche se questo potrebbe essere dovuto al fatto che il numero di uomini sia decisamente minore rispetto alle donne e che i pochi soggetti maschi considerati siano affetti da una neoplasia più aggressiva (o probabilmente anche per motivi biologici).

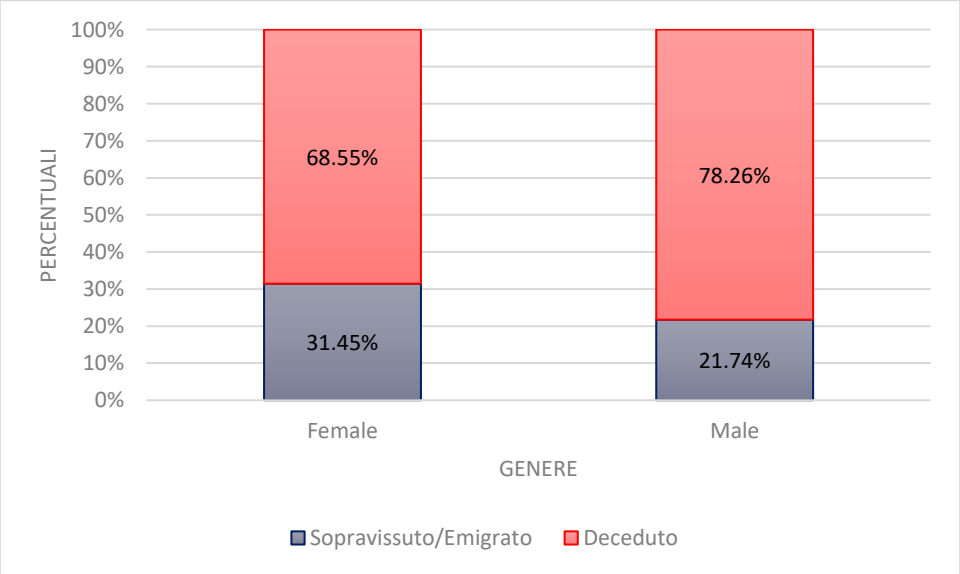

Grafico 4.1 – Distribuzione percentuale della variabile Status per Gender.

Analizzando la distribuzione della variabile età (**Age**), si osserva come vi sia una maggiore frequenza di soggetti che appartengono alle classi di età più avanzate e un numero relativamente basso di individui sotto i 35 anni. Per questo motivo, in una fase più avanzata dell'analisi dove si costruiranno i vari modelli statistici menzionati nel capitolo precedente, la variabile verrà resa categoriale e le sue modalità accorpate in classi (aventi un'ampiezza diversa) per avere un numero di individui per classe maggiore. L'analisi congiunta della variabile risposta con l'età evidenzia un aumento dei decessi causati dalla neoplasia per le fasce più anziane, indipendentemente dal genere. Per le fasce di età più giovani si dispone di una quantità scarsa di osservazioni e dal grafico si osserva come vi sia un sostanziale numero di decessi anche per le fasce più giovani.

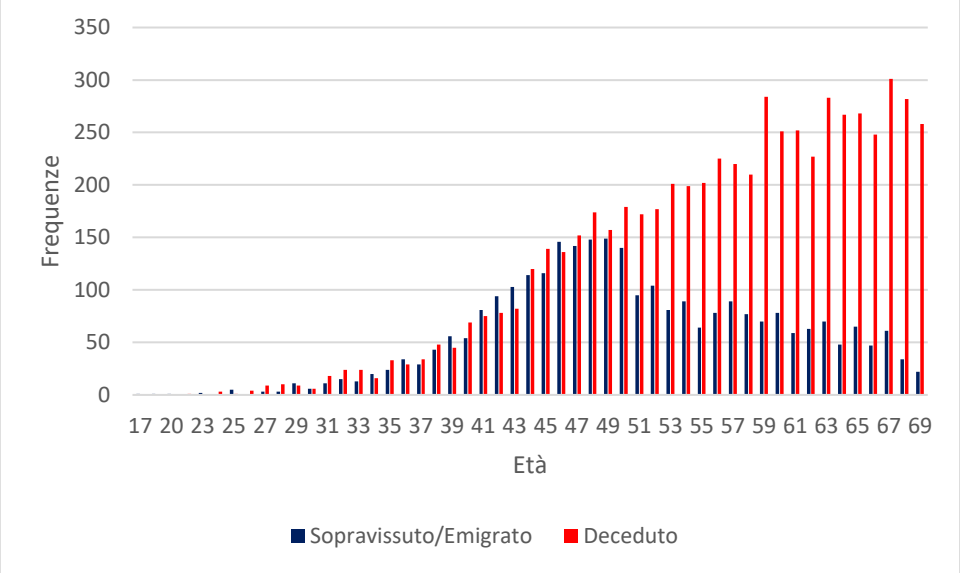

Grafico 4.2 – Distribuzione della variabile Age per Status.

Studiando la distribuzione della variabile categoriale **Localization**, si osserva che la maggior parte degli individui che costituiscono il campione sono affetti da cancro al seno localizzato lateralmente (circa il 33%). Analizzando ulteriormente la distribuzione delle variabili inerenti alle caratteristiche tumorali, quali lo stadio del tumore (**Stadium**) e la tipologia di metastasi (**Metastasis**), si osserva che il 14% dei soggetti presenta un tumore al seno di stadio avanzato (III e IV) e la

maggior parte degli individui presentano metastasi localizzate in prossimità della sede del tumore primitivo. Analizzando congiuntamente queste due variabili, si osserva come i soggetti affetti da tumori di stadio I e II presentino metastasi localizzate nella sede del tumore originario e nelle regioni circostanti, rispettivamente. Gli individui affetti da tumori di stadio più avanzato, invece, presentano metastasi a distanza. Infine, è possibile osservare come la mortalità sia decisamente più elevata per gli individui affetti da tumori avanzati (*Grafico 4.3*).

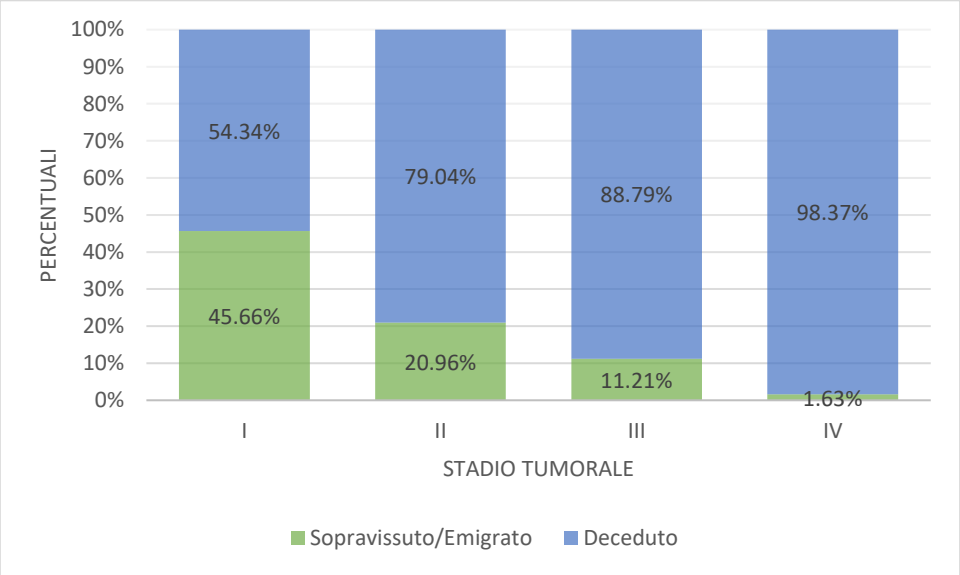

Grafico 4.3 – Distribuzione percentuale della variabile Stadium per Status.

4.4 Analisi della sopravvivenza

L'analisi della sopravvivenza è senz'altro una tematica importante e diffusa in molte discipline, che spaziano dall'ambito medico e biologico sino a quello ingegneristico, dove trova sempre più larga applicazione. Un primo passo per iniziare questo studio è rappresentato dalla stima delle curve di sopravvivenza, le quali sono strumenti molto utili sia per eseguire un'analisi preliminare dei dati sia per calcolare alcune quantità derivate dai modelli di regressione (come il tempo mediano di sopravvivenza). Vi sono diverse procedure che consentono la stima delle curve di sopravvivenza e in questo lavoro si è deciso di applicare il metodo di **Kaplan-Meier**.

Lo stimatore di Kaplan-Meier (*KM*) consiste in uno dei metodi più utilizzati per la stima della funzione di sopravvivenza, soprattutto in ambito biomedico. Nei paragrafi a seguire tale stimatore verrà indicato come $\tilde{S}(t)$. Per capire al meglio il suo funzionamento, si riprenda la definizione di funzione di sopravvivenza, definita al Capitolo 2 (definizione (2.2)), ovvero si intenda $S(t)$ come la probabilità che il soggetto non abbia sperimentato l'evento di interesse fino al tempo t.

Quando non si è in presenza di dati censurati, $\tilde{S}(t)$ sarà semplicemente uguale alla proporzione di individui, al tempo t, per i quali deve ancora verificarsi l'evento. Tuttavia questa prima situazione si verifica di rado, poiché nei contesti dell'analisi di sopravvivenza si è soliti lavorare con dati censurati. Nell'ipotesi di singola censura a destra, ovvero nella situazione in cui i casi sono censurati nel medesimo istante temporale c e tutti gli eventi osservati si verificano entro questo periodo di tempo, allora $\tilde{S}(t)$ viene calcolata come nel caso precedente, $\forall\, t \leq c$. Per $\forall\, t > c$, $\tilde{S}(t) = \tilde{S}(c) \forall\, t > c$.

Il secondo contesto possibile e allo stesso tempo più comune rispetto al primo, prevede che alcuni casi siano censurati in diversi istanti temporali dello studio e quindi, a differenza del caso precedente, vi siano alcuni soggetti che sperimenteranno l'evento anche dopo gli istanti di censura. In questo contesto, se si decidesse di calcolare la proporzione di soggetti che devono ancora sperimentare l'evento di interesse per un dato istante temporale t (come nel caso

precedente), si otterrebbe una stima distorta verso il basso, in quanto i casi censurati prima del tempo t potrebbero aver sperimentato l'evento. Per risolvere questa problematica si procede a suddividere l'asse del tempo in k punti temporali: $0 < t_1 < t_2 < \cdots < t_j < \cdots < t_k < \infty$, andando a formare $k + 1$ intervalli in corrispondenza dei quali è avvenuto almeno un evento, del tipo: $I_1 = [0, t_1)$, $I_j = [t_j, t_{j-1})$ e $I_k = [t_k, \infty)$.

All'inizio di ogni intervallo I_j vi sono n_j individui che sono a rischio di subire l'evento di interesse, ovvero soggetti che non lo hanno ancora sperimentato o non sono stati censurati nell'intervallo precedente. Se alcuni soggetti sono censurati esattamente al tempo t_j vengono comunque considerati soggetti a rischio di subire l'evento nell'intervallo I_j. Sia q_j il numero di individui che sperimentano l'evento al tempo t_j (nel nostro caso saranno coloro che decedono per il cancro al seno), allora lo stimatore *KM* viene definito come:

$$\tilde{S}(t) = \prod_{j:t_j \leq t} \left[1 - \frac{q_j}{n_j}\right], \tag{4.1}$$

per $t_1 \leq t \leq t_k$. In altri termini, per un dato istante temporale t, si considerano tutti gli eventi che si sono verificati negli intervalli precedenti a t e, per ognuno di questi, si calcola la quantità definita tra parentesi, la quale può essere interpretata come una stima della probabilità condizionata di sopravvivere fino al tempo t_{j+1}, dato che si è sopravvissuti fino al tempo t_j. Infine si procede a moltiplicare tra loro tutte queste probabilità condizionate, ottenendo così la stima della funzione di sopravvivenza al tempo t. Per $t \leq t_1$, $\tilde{S}(t) = 1$, mentre per $t \geq t_k$ la definizione di $\tilde{S}(t)$ dipende dalla configurazione delle osservazioni censurate. Infatti, quando non sono presenti censure oltre l'istante temporale t_k, $\tilde{S}(t) = 0$ per $\forall\, t > t_k$, viceversa, quando vi sono osservazioni censurate oltre t_k, $\tilde{S}(t)$ è positiva $\forall\, t > t_k$ (Allison, 1995, p. 29-31).

Si procede a prendere in considerazione tutti i soggetti affetti da neoplasia mammaria che costituiscono il dataset presentato in precedenza. Ogni individuo viene seguito per un periodo pari alla differenza, espressa in anni, dell'anno di uscita ed entrata in studio (ovvero l'anno in cui viene diagnosticato il tumore al seno e l'anno del decesso/uscita dallo studio, rispettivamente). Quindi, come periodo di tempo si procede a considerare la differenza: **YEAR_STATUS** - **YEAR_DIAGNOSIS**. Sulla base di questi pretesti, viene riportata di seguito la curva di sopravvivenza in relazione alla mortalità causata da cancro al seno, per gli individui appartenenti al dataset. In questo caso sono considerati come casi censurati tutti i soggetti che non sono morti a causa della neoplasia durante il periodo di osservazione (pari a 26 anni), ovvero coloro che sopravvivono fino alla fine dello studio o che escono dall'osservazione per ragioni diverse dalla morte cancro-correlata (in questo caso per emigrazione). Nel dataset questi soggetti hanno il valore della variabile **Status** pari a 0.

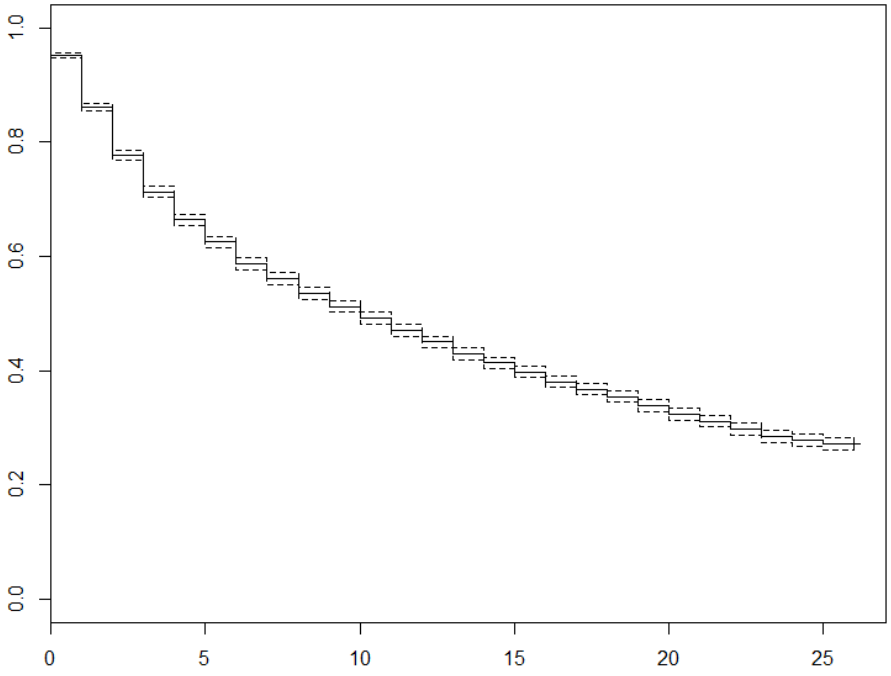

Grafico 4.4 – Curva di sopravvivenza dei soggetti dello studio, stimata con metodo KM.

Uno degli indici più utili per descrivere brevemente il fenomeno di interesse è il tempo mediano di sopravvivenza. Tale indicatore rappresenta l'istante temporale in cui il 50% dei soggetti in studio è ancora in vita ed è pari a 10.17 anni (3714 giorni) con un intervallo di confidenza al 95% pari a [9.68 – 10.61].

Sulla base di quanto detto, il cancro al seno è una neoplasia che colpisce in modo particolare le donne, ma è possibile che questa malattia insorga anche nell'uomo. Per tale motivo vengono riportate le curve di sopravvivenza distinte per genere e, in una fase immediatamente successiva, si procederà al loro confronto tramite il Test dei ranghi logaritmici.

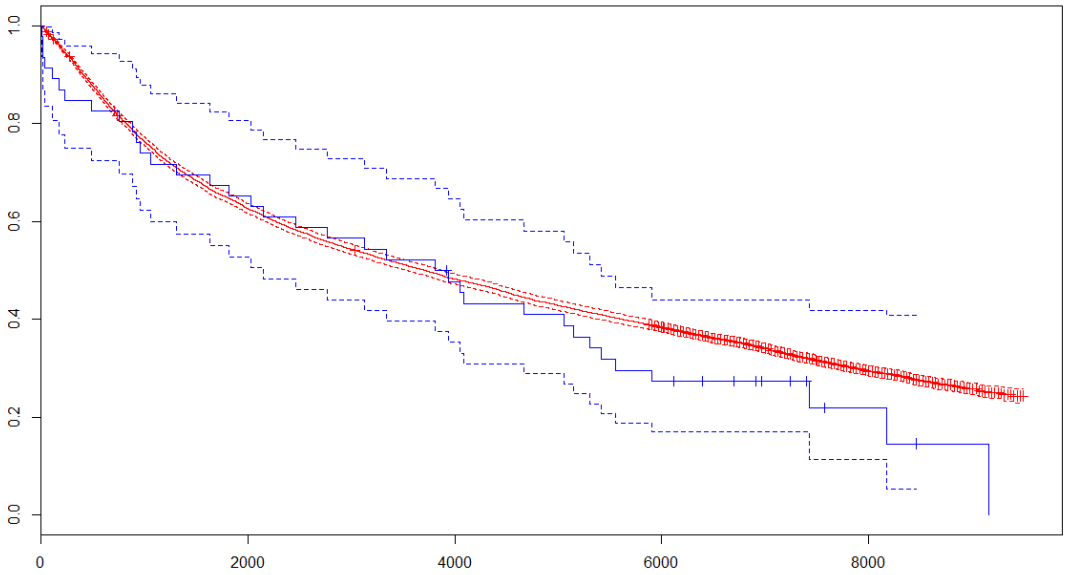

Grafico 4.5 – Curve di sopravvivenza dei pazienti di sesso maschile (blu) e femminile (rosso), stimate con metodo KM.

Il Test dei ranghi logaritmici (*o Log-Rank Test*) è uno strumento che consente di confrontare tra loro due o più curve di sopravvivenza. Questo risulta essere di particolare rilevanza quando si desidera confrontare la sopravvivenza, per esempio, di due gruppi di pazienti: il primo che riceve un nuovo farmaco per il trattamento di una certa tipologia tumorale, mentre al secondo gruppo di pazienti viene somministrato il trattamento standard per lo stesso tipo di tumore. Questo è un test non parametrico (che è caratterizzato da una procedura simile, ma non identica al χ^2) che si basa su una quantità, al numeratore, che somma le discrepanze tra il numero di eventi osservati e il numero di eventi attesi per di ogni istante temporale, in corrispondenza dei quali è avvenuto almeno un evento. Questa operazione viene eseguita per ogni gruppo per il quale avviene il confronto delle curve di sopravvivenza. Per il gruppo i-esimo, con $i = 1, \dots, L$ e per r periodi di tempo, tale quantità viene definita come:

$$D = \sum_{j=1}^{r} (m_{ij} - e_{ij}). \tag{4.2}$$

La quantità m_{ij} rappresenta il numero di eventi che si verificano per il gruppo i-esimo nell'istante temporale j e, nel caso di questo lavoro, sono le morti di cancro al seno avvenute nei diversi periodi di tempo. Il secondo elemento e_{ij}, invece, è il numero atteso di morti per il gruppo i al tempo j. Quest'ultima quantità viene calcolata come $\frac{n_{ij}m_j}{n_j}$, dove n_{ij} è il numero di soggetti a rischio di sperimentare l'evento di interesse prima del tempo j, nel gruppo i-esimo; m_j è il numero totale di eventi che si verificano al tempo j per tutti i gruppi mentre n_j è il numero totale di casi a rischio di sperimentare l'evento prima dell'istante temporale j. Questa formula è utilizzata nel contesto in cui si abbia a che fare con una tabella di contingenza a entrata multipla, sotto ipotesi di indipendenza. L'ipotesi nulla H_0 sottostante il test prevede che le diverse curve di sopravvivenza provengano dalla stessa popolazione. La statistica-test definita al (4.2) viene successivamente divisa per la varianza: $Var(D)$. Il risultato ottenuto è un χ^2 e il test risulterà essere

statisticamente significativo, ovvero verrà rifiutata l'ipotesi nulla, nel caso in cui il valore del Chi-quadro ricavato risulterà essere maggiore di un χ^2 avente $L-1$ gradi di libertà, con L pari al numero di gruppi che si intende confrontare.

Nel nostro caso disponiamo di un numero di gruppi pari a 2, in quanto l'intenzione è quella di confrontare le curve di sopravvivenza dei due generi di pazienti. Il risultato del Test dei ranghi logaritmici è pari a 1.4 e, confrontandolo con un χ_1^2 (3.84), viene accettata l'ipotesi nulla H_0 (p-value: 0.229). Pertanto le due curve possono essere considerate statisticamente equivalenti (Allison, 1995), (Dalgaard, 2008).

Nel corso delle analisi preliminari, sono state confrontate le curve di sopravvivenza relative alle modalità di altre variabili esplicative, oltre a quella rappresentante il genere dei pazienti, quali: **Localization**, **Metastasis**, **Stadium** e **Intervention**. Lo scopo consiste nel valutare se la sopravvivenza dei diversi gruppi di pazienti, che si identificano nei diversi livelli delle variabili, potesse essere la stessa o se fossero presenti delle differenze, magari anche interessanti. Dai risultati è emerso che i soggetti che sono affetti da una neoplasia che coinvolge tutto il seno (modalità "Both breast" della variabile **Localization**) hanno una probabilità di sopravvivere più bassa rispetto agli altri gruppi di pazienti. Infatti il Test dei ranghi logaritmici ha portato al rifiuto dell'ipotesi di uguaglianza delle quattro diverse curve (p-value: 0). Il medesimo test ha portato al rifiuto anche in tutti gli altri casi (tutti i p-value sono prossimi allo 0) e, dallo studio delle diverse curve, si è osservato come i pazienti che presentano delle metastasi a distanza, con uno stadio del tumore avanzato (IV stadio) e coloro che non ricevono alcuna operazione chirurgica (modalità "No surgery" della variabile **Surgery1**) abbiano una probabilità di sopravvivere molto più bassa.

4.5 Preparazione del dataset

Il dataset a nostra disposizione presenta, come affermato in precedenza, 9041 osservazioni: per ogni soggetto, rappresentato da un identificativo specifico, vengono riportate una serie di informazioni riguardanti le proprie caratteristiche socio-demografiche e della malattia. Il formato con cui i dati sono riportati non risulta essere il più adatto per l'analisi che si desidera condurre. Nell'espressione (2.30) del Capitolo 2 è stata definita la log-verosimiglianza per il modello Poisson applicato in un contesto di analisi di sopravvivenza. Si è dimostrato come quest'espressione non differisca da quella di un modello esponenziale a rischi proporzionali e costanti a tratti con osservazioni censurate a destra, se non per una costante dipendente dai dati, la quale può essere tranquillamente trascurata. In tale formula si osserva come il contributo individuale alla log-verosimiglianza si presenti come una somma di un certo numero di termini, ognuno specifico per un dato intervallo. Più precisamente, ogni individuo appartenente allo studio determina il proprio contributo in ciascun intervallo che visita, da quello relativo alla propria entrata in studio fino all'intervallo di tempo in cui sperimenta l'evento di interesse o viene censurato, indicato come $j(i)$: in tutti gli intervalli precedenti a $j(i)$, l'individuo non sperimenta l'evento e quindi il contributo è rappresentato solamente dal prodotto del tempo di esposizione per il rischio nello specifico intervallo j, in formule: $t_{ij}\lambda_{ij}$.

Occorre modificare il formato del dataset affinché possa essere implementato il modello di Poisson presentato al Capitolo 2. A tale scopo, si suddivide l'osservazione di ciascun soggetto i-esimo in $j(i)$ sotto-osservazioni, per $j = 1, \dots, J$. I sotto-intervalli sono costruiti sulla scala temporale "Età", usando la variabile **etanew**. Essi hanno tutti ampiezza quinquennale, tranne il primo. Gli intervalli in questione sono: [17-35), [35-40), [40-45), [45-50), [50-55), [55-60), [60-65), [65-70), [70-75), [75-80), [80-85), [85-90), creati codificando la variabile **Age** come categoriale, raggruppandone le diverse modalità. L'ampiezza del primo intervallo è pari a [17-35) e ciò è dovuto alla necessità di disporre di un numero maggiore di osservazioni, che non si sarebbero avute considerando più intervalli di ampiezza pari a cinque nei primi anni. Per ognuno di questi J intervalli, si

considera il contributo di ciascun individuo i-esimo. Per avere un'idea più chiara di quello di cui si sta parlando, si consideri per esempio il paziente avente l'identificativo (variabile **ID**) 345.

ID	Status	YEAR_STATUS	YEAR_DIAGNOSIS	Age	etaclass
345	0	1991	1968	66	[65-70)

ID	Status	YEAR_STATUS	YEAR_DIAGNOSIS	Age	etanew	eta.ent.int	eta.usc.int	anno.ent.int
345	0	1991	1968	66	[65-70)	66	70	1968
345	0	1991	1968	70	[70-75)	70	75	1972
345	0	1991	1968	75	[75-80)	75	80	1977
345	0	1991	1968	80	[80-85)	80	85	1982
345	0	1991	1968	85	[85-90)	85	89	1987

Tabella 4.1 –Suddivisione del periodo di follow-up del soggetto avente identificativo 345.

Tale soggetto viene tenuto sotto osservazione per 23 anni e non risulta morire a causa di tumore al seno, poiché il valore della variabile **Status**, per questo paziente, assume valore 0. L'idea è quella di creare differenti sotto-osservazioni per l'individuo, in modo tale che ciascuna di queste rientri in un intervallo specifico.

Inizialmente si procede alla creazione di una nuova variabile, che nel contesto verrà chiamata: **anni**. Essa rappresenta la differenza tra l'anno in cui il soggetto decede/sopravvive per il tumore al seno (**YEAR_STATUS**) e l'anno in cui gli viene diagnosticata la neoplasia (**YEAR_DIAGNOSIS**), e viene quindi intesa come il periodo in cui l'individuo è esposto all'evento morte, espresso in termini di anni o più semplicemente come periodo di follow-up dell'individuo. Per i valori della variabile anni che vanno da 5 a 26, la procedura di sistemazione del dataset consta di due fasi: nella prima si procede a duplicare le osservazioni degli individui, al fine di ottenere una coppia di osservazioni identiche per ogni soggetto. Per ognuna di esse vengono create due nuove variabili, che rappresentano l'età del soggetto all' entrata dell'intervallo (**eta.entrata.intervallo**) e l'età di uscita dallo stesso (**eta.uscita.intervallo**). Per

ogni coppia di osservazioni si procede a modificare l'età in uscita della prima, assegnando l'estremo superiore dell'intervallo *j* più vicino a quel valore, l'età in entrata della seconda e i valori della variabile Status. La seconda fase della sistemazione prevede di applicare la funzione "timeSplitter", disponibile nella libreria "Greg" di R, alla seconda metà del dataset (le osservazioni duplicate dei soggetti). Essa permette di suddividere il periodo di follow-up di ogni soggetto in più intervalli temporali, ciascuno delimitato da un periodo di inizio e di fine (definiti nell'output di R come **Start_time** e **Stop_time**, rispettivamente). L'operazione di "splitting" viene eseguita sulla base della variabile **anni**, ovvero il periodo di esposizione di ciascun soggetto viene diviso in più intervalli, la cui ampiezza è stata fissata pari a cinque anni. Quindi, per ogni individuo, si ottengono più sotto-osservazioni, in cui quasi tutti i valori delle variabili esplicative rimangono fissi, fatta eccezione per la variabile **Status**, la quale assume valore 0 in tutti gli intervalli precedenti all'ultimo intervallo effettivamente visitato dal soggetto, e valori 0 o 1 nell'ultimo intervallo, a seconda che l'individuo abbia sperimentato l'evento di interesse o sia stato censurato, rispettivamente. Infine è stata creata la variabile **anno.entrata.intervallo**, la quale rappresenta l'anno in cui il soggetto entra nell'intervallo j-esimo e sono state aggiornate i valori relativi all'età in entrata e in uscita dei soggetti per ciascuna "sotto-osservazione".

Per i restanti valori della variabile **anni** (0, 1, 2, 3 e 4) è stata eseguita una procedura leggermente differente poiché il metodo specificato in precedenza non portava ai risultati desiderati. Anche in questi casi, tuttavia, si è proceduto a duplicare le osservazioni di ciascun individuo e alla creazione delle stesse variabili, ragionando in maniera differente a seconda del periodo di follow-up. Per esempio, per coloro che sono caratterizzati da un periodo di esposizione pari a 0, è bastato assegnare lo stesso valore alle variabili **eta.entrata.intervallo** ed **eta.uscita.intervallo**, l'anno di entrata nell'intervallo coincideva con l'anno in cui è stato diagnosticato il tumore mentre la variabile **Status** è rimasta invariata.

4.6 Stima del modello di Poisson e selezione del modello

Dopo aver conseguito la sistemazione del dataset, si procede alla stima di un modello di Poisson, del tipo:

$$log\mu_{ij} = logt_{ij} + x_i'\beta \ + \sum_{j=1}^{J} I_j\beta_{0j}$$

Lo scopo dell'analisi consiste nello studio della mortalità causata dal cancro al seno in un campione della popolazione norvegese, in particolare si desidera scoprire se tale mortalità aumenta in maniera uniforme su tutte le fasce di età o più rapidamente in alcune di queste. Un obiettivo secondario risiede nell' approfondire la relazione esistente tra le morti di cancro e le caratteristiche socio-demografiche e della malattia dei soggetti. Sulla base dei concetti visti nel capitolo precedente, si tratterà la risposta d_{ij} come realizzazione di una variabile casuale avente distribuzione Poisson, di media $\mu_{ij} = t_{ij}\lambda_{ij}$. Il primo termine rappresenta il periodo di esposizione del soggetto i nell'intervallo j, mentre λ_{ij} è il rischio di sperimentare l'evento morte in questo intervallo per l'individuo i-esimo. Quindi il numero di eventi attesi μ_{ij} per ciascun paziente è il prodotto del tasso di mortalità per il tempo di esposizione. Nel modello sono state prese in considerazione come variabili esplicative, $x_i'\beta$: **Gender**, **Localization**, **Metastasis**, **Stadium** e **anno.entrata.intervallo**, quest'ultima opportunamente trasformata in una variabile categoriale, avendone accorpato le modalità, dato che se si fosse lasciata intesa come numerica, si avrebbe avuto un numero decisamente più elevato di predittori. L'offset definito nel modello lineare generalizzato è $\log(t_{ij})$, ossia il logaritmo della differenza di età del soggetto i-esimo tra la propria uscita ed entrata nell'intervallo temporale j. Infine si è proceduto a inserire nel modello la variabile categoriale **etanew**, la quale rappresenta la fascia di età a cui appartiene ciascun soggetto in ogni specifico intervallo j che visita nel corso della sua permanenza nello studio. Le modalità di tale variabile sono: [17-35), [35-40), [40-45), [45-50), [50-55), [55-60), [60-65), [65-70), [70-75), [75-80), [80-85), [85-90) e corrispondono agli intervalli in cui i pazienti sono a rischio di sperimentare l'evento di interesse. Pertanto, ogni

coefficiente associato a queste variabili, β_{0j} rappresenta il rischio di morte costante relativo all'intervallo j-esimo.

Viene presentata di seguito una sintesi del modello che include tutte le variabili esplicative dette in precedenza:

Coefficiente	Stima	Std. Error	z-value	Pr(>|z|)
Intercetta	-0,4934	0,58316	-0,846	0,397514
etanew[35-40)	-0,4584	0,19589	-2,34	0,019288 *
etanew[40-45)	-0,7491	0,17324	-4,324	1,53e-05 ***
etanew[45-50)	-0,8463	0,16026	-5,281	1,29e-07 ***
etanew[50-55)	-0,6905	0,15516	-4,45	8,59e-06 ***
etanew[55-60)	-0,5261	0,15387	-3,419	0,000628 ***
etanew[60-65)	-0,4081	0,15305	-2,666	0,007670 **
etanew[65-70)	-0,3132	0,15256	-2,053	0,040097 *
etanew[70-75)	-0,1777	0,15586	-1,14	0,254388
etanew[75-80)	0,427	0,16094	2,653	0,007973 **
etanew[80-85)	0,79226	0,17392	4,555	5,23e-06 ***
etanew[85-90)	1,19225	0,20829	5,724	1,04e-08 ***
GenderMale	0,39138	0,21504	1,82	0,068762
LocalizationCentral	-0,4505	0,12421	-3,627	0,000287 ***
LocalizationLateral	-0,5048	0,11861	-4,256	2,08e-05 ***
LocalizationMedian	-0,2922	0,12209	-2,393	0,016709 *
LocalizationWhole breast	0,09671	0,13791	0,701	0,483157
MetastasisLocalized	-1,6036	0,55925	-2,867	0,004138 **
MetastasisRegional	-1,1557	0,50532	-2,287	0,022197 *
StadiumII	0,38543	0,24857	1,551	0,120999
StadiumIII	0,73646	0,24776	2,972	0,002955 **
StadiumIV	0,22993	0,56245	0,409	0,682684
anno.entrata.intervallo[1970-1975)	-0,2098	0,04407	-4,761	1,92e-06 ***
anno.entrata.intervallo[1975-1980)	-0,5576	0,05497	-10,143	<2e-16 ***
anno.entrata.intervallo[1980-1985)	-0,8733	0,06515	-13,404	<2e-16 ***
anno.entrata.intervallo[1985-1990)	-1,1355	0,08112	-13,997	<2e-16 ***

Tabella 4.2 –Summary del modello di Poisson per dati di sopravvivenza applicato al dataset reale.

Si procede all'interpretazione delle stime dei parametri. L'intercetta del modello include i soggetti aventi un'età compresa tra i 17 e 34 anni, di genere femminile, con il cancro localizzato su entrambi i seni, con metastasi localizzate a distanza dalla sede del tumore primitivo e aventi un tumore di fase iniziale (I stadio); tutte le interpretazioni che seguiranno verranno espresse considerando tali individui come gruppo di riferimento. Quasi tutte le stime dei parametri associati alle modalità della variabile **etanew**, e quindi agli intervalli, sono risultate significative, fatta eccezione per etanew[70-75). La significatività di tali coefficienti ci porta a dedurre che il rischio di morte per cancro al seno non sia uguale per tutte le fasce di età, di conseguenza sembra essere corretto ipotizzare un rischio non costante per l'età dei pazienti. In particolare si osserva un calo della mortalità, che prosegue fino ai 45-50 anni per poi aumentare, in modo sempre più considerevole, nelle fasce di età più anziane (si veda Grafico 4.6): l'effetto moltiplicativo per i soggetti che cadono nell'intervallo di età [45-50) è pari a $\exp\{-0.8463\} = 0.4290$, evidenziando un calo della mortalità del 57% comparato con il gruppo di riferimento, al netto delle altre variabili esplicative. Come ci si poteva aspettare, la fascia di età che presenta un rischio maggiore di morte per tumore al seno corrisponde a quella dei soggetti aventi più di 85 anni, con fattore moltiplicativo pari a $\exp\{1.19225\} = 3.2945$, che porta a un rischio decisamente più elevato rispetto ai soggetti con le caratteristiche del gruppo di base. Spostando l'attenzione sui predittori che si è deciso di inserire nel modello, è possibile osservare come il fatto di appartenere al genere maschile, di avere un tumore localizzato su tutto il seno (modalità: "Whole breast" della variabile **Localization**) e di avere un tumore di stadio avanzato ("Stadium III"), aumenti la mortalità dei pazienti, rispetto ai soggetti appartenenti al gruppo di base, di un fattore moltiplicativo pari a $\exp\{0.39138 + 0.09671 + 0.73646\} = 3.4026$. Eseguendo il test del log-rapporto di verosimiglianza, è possibile concludere che le variabili esplicative inserite nel modello di Poisson riescano a descrivere in maniera adeguata la mortalità legata alla neoplasia per la popolazione considerata.

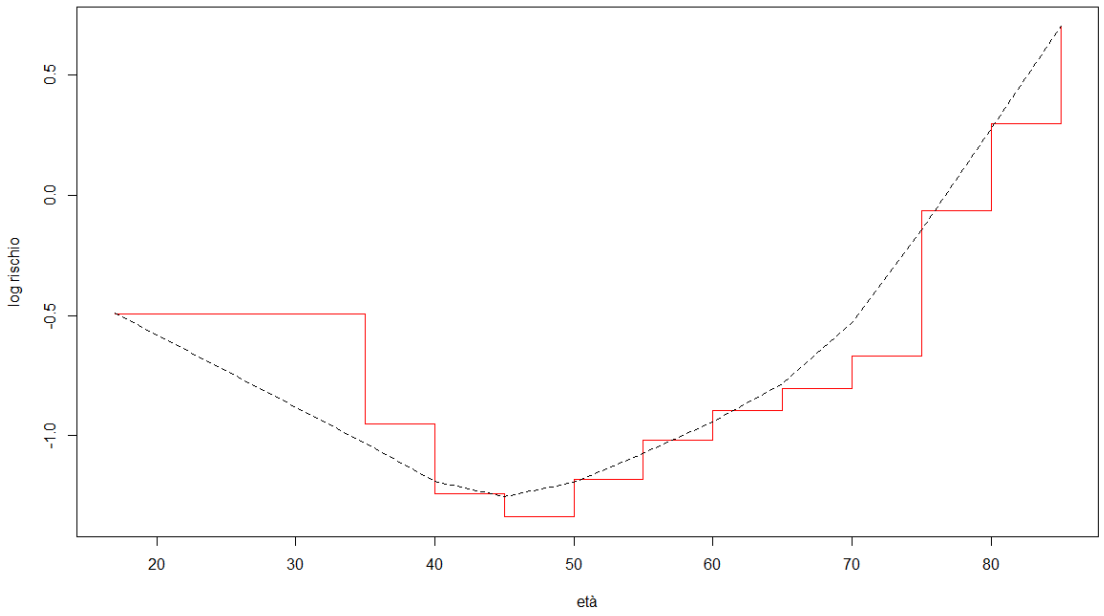

Grafico 4.6 – Funzione a gradini rappresentante il rischio di morte (su scala logaritmica) per cancro al seno al variare delle classi di età, a cui viene sovrapposta una spline di lisciamento con parametro di regolazione pari a 0.4.

L'analisi del modello di Poisson ottenuto può essere ulteriormente estesa al calcolo dei rischi cumulati e, di conseguenza, alla stima delle probabilità di sopravvivenza. Infatti, potrebbe essere di diretto interesse valutare se, tra le diverse classi di età a cui appartengono i soggetti, vi sia una diversa probabilità di sopravvivere al tumore. Per una rappresentazione più compatta e riassuntiva, i risultati dei vari calcoli eseguiti vengono esposti nella tabella sottostante (Tabella 4.3):

Fascia di età	Ampiezza intervallo	Log-Hazard	Hazard	Cum. Hazard	Survival probabilities
[17-35)	18/73	-0.49339	0.610553106	0.150547341	0.860237004
[35-40)	5/73	-0.95175	0.386064818	0.176990137	0.837788047
[40-45)	5/73	-1.24247	0.28867032	0.196761851	0.821386222
[45-50)	5/73	-1.33968	0.261929472	0.214702226	0.806781643
[50-55)	5/73	-1.18386	0.306094928	0.235667632	0.790043215
[55-60)	5/73	-1.01948	0.360782498	0.260378762	0.770759596
[60-65)	5/73	-0.90147	0.405972441	0.288185093	0.749622829
[65-70)	5/73	-0.80656	0.446391013	0.31875982	0.727050151
[70-75)	5/73	-0.67104	0.511176677	0.353771922	0.702035067
[75-80)	5/73	-0.06639	0.935765844	0.417865472	0.658450802
[80-85)	5/73	0.29887	1.348334329	0.510217139	0.600365202
[85-90)	5/73	0.69886	2.011458337	0.647988258	0.523097055

Tabella 4.3 – Stima delle probabilità di sopravvivenza per le differenti fasce di età dei soggetti.

Le prime due colonne rappresentano, rispettivamente, le diverse fasce di età a cui appartengono i soggetti nei vari intervalli j che visitano nel corso dello studio, e le ampiezze degli intervalli temporali, espresse in anni. Il rischio λ_{ij} rappresenta il tasso di morte per cancro per anni-persona di esposizione. La terza, quarta e quinta colonna sono il logaritmo del rischio $\log(\lambda_{ij})$, ottenuto come somma del valore dell'intercetta con i diversi coefficienti associati alle variabili dell'età, il rischio λ_{ij}, ricavato attraverso la trasformazione esponenziale dei precedenti valori e il rischio cumulato $H_i(t)$, calcolato moltiplicando l'ampiezza di ciascun intervallo j per il rischio associato e sommando i termini man mano che si procede con gli intervalli. L'ultima colonna rappresenta la funzione di sopravvivenza, i cui valori vengono stimati applicando la trasformazione esponenziale ai valori del rischio cumulato cambiati di segno. Da questi è possibile ricavarsi la probabilità di sopravvivere al tumore per ciascuna fascia di età. Per esempio la probabilità di sopravvivenza per i soggetti aventi più di 85 anni è del 52,3%.

Il modello individuato in precedenza sembra spiegare in maniera adeguata la relazione esistente tra la mortalità causata dal tumore al seno e le caratteristiche socio-demografiche dei pazienti, in particolare con l'età del soggetto. Tuttavia si preferisce essere sicuri che quello individuato sia il miglior modello possibile per la rappresentazione del fenomeno di interesse. La domanda che spesso ci si pone è

se le variabili inserite nel modello siano in grado di spiegare bene l'outcome di interesse oppure se è necessario procedere alla loro rimozione, dato che il loro apporto è solamente confusionario. Per questo motivo si è deciso di proseguire con una procedura di selezione di variabili, la regressione stepwise, che si basa sul criterio di Akaike, meglio noto come AIC (*Akaike's Information Criterion*). A ogni passo del processo viene inserito o rimosso quel predittore che maggiormente migliora l'adattamento del modello, fino a quando nessun altro termine risulta essere necessario in termini di abbassamento dell'AIC. Il modello risultante è del tutto identico a quello ottenuto in precedenza e per questo si ricorre a un'alternativa, la quale consiste nell'applicazione di un test χ^2 a ogni termine che caratterizza il modello completo, attraverso il comando *drop1*. Di seguito viene riportato l'output dell'operazione svolta:

	Df	**Deviance**	**AIC**	**LRT**	**Pr(>Chi)**
<none>			19315		
etanew	11	12806	19314	421,2	< 2,2e-16 ***
Gender	1	12388	19316	2,92	0,08729
Localization	4	12457	19379	71,94	8,840e-15 ***
Metastasis	2	12392	19318	7,31	0,02586 *
Stadium	3	12419	19343	34,08	1,906e-07 ***
anno.entrata.intervallo	4	12702	19624	316,7	< 2,2e-16 ***

Tabella 4.4 – Risultati del test χ^2 applicato al modello di Poisson per dati di sopravvivenza .

Ciascuna delle righe che costituiscono l'output ci informa se il modello senza quel determinato predittore risulti essere significativamente differente dal modello completo. Dal risultato ottenuto, si osserva come l'unica variabile a non essere significativa sia quella inerente al genere del soggetto: **Gender**, dato che il p-value associato è pari a 0.08729. Sulla base di questo risultato, viene scelto il modello più semplice. Appare quindi evidente che il fatto di appartenere al genere maschile o femminile non influisce in alcun modo sulla mortalità causata dalla neoplasia mammaria.

4.7 Applicazione dello stimatore non distorto in mediana

Il valore aggiunto che si vuole presentare in questo lavoro consiste nella stima del modello di Poisson visto in precedenza attraverso la libreria brglm2. Questo pacchetto di R permette di ottenere stime non distorte in mediana dei parametri di interesse attraverso una modifica dell'equazione di verosimiglianza (si veda Capitolo 3). Lo scopo di quest'applicazione risiede nell'identificare eventuali differenze che potrebbero sorgere dall'utilizzo dei due differenti metodi di stima. Per stimare il modello di interesse, viene applicato il comando **brglm_fit**, il quale non richiede sostanziali cambiamenti dai soliti comandi usati per la stima dei Modelli Lineari Generalizzati. L'unico accorgimento consiste nello specificare che le stime dei parametri di interesse siano non distorte in mediana, tramite l'argomento **type = "AS_median"**, dato che l'argomento di default permette di ottenere una non distorsione in media dei predittori. Per un confronto diretto, si decide di cominciare con il modello completo, il quale include anche la variabile **Gender**. Osservando le stime dei parametri, si nota come ci siano delle lievi differenze rispetto a quelle del modello di Poisson stimato in precedenza. Le stime dei parametri inerenti le variabili **etanew** e **Localization** risultano essere leggermente più piccole ed è possibile osservare lo stesso cambiamento per la variabile **Metastasis**, anche se questa differenza risulta essere decisamente più elevata. La maggior parte delle variabili che riguardano lo stadio tumorale presentano delle stime leggermente più grandi rispetto a quelle individuate con lo stimatore classico, mentre per l'esplicativa *StadiumIV* si osserva un calo molto più marcato, dato che si passa da una stima pari a 0.22993 a 0.18460, indicando un rischio di morte minore per i soggetti che sono affetti da un tumore di fase IV. La stima del parametro associato alla variabile **Gender** non differisce di molto, mentre le stime per i predittori inerenti all'anno di entrata dei soggetti negli intervalli subiscono delle leggere modifiche, le quali non sono uniformi. Infatti, la stima per *anno.entrata.intervallo[1970-1975)* diminuisce leggermente mentre le stime per le restanti tre variabili aumentano. Nonostante qualche cambiamento nella stima dei parametri, l'interpretazione del modello rimane sostanzialmente immutata. Infatti, è possibile osservare un calo del rischio di morte per cancro al seno fino ai 50 anni seguito da un successivo e continuo aumento nelle fasce di

età più anziane, proprio come risultato in precedenza. Infine è stato eseguito un test rapporto di verosimiglianza tra il modello completo e il modello senza la variabile inerente al genere del paziente, al fine di individuare il miglior modello che spieghi il fenomeno di interesse. Nuovamente la variabile **Gender** non risulta essere significativa. È possibile interpolare lo stesso identico modello anche attraverso il comando **glm**, con l'inserimento di opportuni codici all'interno dello specifico comando. Si decide di riportare l'output che ne consegue in quanto risulta avere un'interpretazione più agevole e familiare allo stesso tempo.

Coefficiente	Stima	Std. Error	z-value	Pr(>|z|)
Intercetta	-0,42594	0,56713	-0,751	0,452624
etanew[35-40)	-0,45903	0,19334	-2,374	0,017585 *
etanew[40-45)	-0,75113	0,17102	-4,392	1,12e-05 ***
etanew[45-50)	-0,84897	0,15818	-5,367	8,00e-08 ***
etanew[50-55)	-0,69339	0,15312	-4,528	5,94e-06 ***
etanew[55-60)	-0,52911	0,15184	-3,485	0,000493 ***
etanew[60-65)	-0,4112	0,15104	-2,722	0,006480 **
etanew[65-70)	-0,31631	0,15055	-2,101	0,0035642 *
etanew[70-75)	-0,1807	0,15383	-1,175	0,24013
etanew[75-80)	0,42376	0,15887	2,667	0,007645 **
etanew[80-85)	0,78929	0,17174	4,596	4,31e-06 ***
etanew[85-90)	1,19136	0,2057	5,792	6,96e-09 ***
GenderMale	0,39872	0,21215	1,879	0,06189
LocalizationCentral	-0,45231	0,12282	-3,683	0,000231 ***
LocalizationLateral	-0,50678	0,11728	-4,321	1,55e-05 ***
LocalizationMedian	-0,294	0,12073	-2,435	0,014883 *
LocalizationWhole breast	0,09537	0,13635	0,699	0,484269
MetastasisLocalized	-1,64949	0,54314	-3,037	0,002390 **
MetastasisRegional	-1,195	0,48868	-2,445	0,014470 *
StadiumII	0,38895	0,24587	1,582	0,113658
StadiumIII	0,74067	0,24508	3,022	0,002510 **
StadiumIV	0,1846	0,54638	0,338	0,735463
anno.entrata.intervallo[1970-1975)	-0,20995	0,04357	-4,818	1,45e-06 ***
anno.entrata.intervallo[1975-1980)	-0,55744	0,05436	-10,254	<2e-16 ***
anno.entrata.intervallo[1980-1985)	-0,87302	0,06443	-13,549	<2e-16 ***
anno.entrata.intervallo[1985-1990)	-1,13493	0,08023	-14,146	<2e-16 ***

Tabella 4.5 – Stime dei parametri non distorti in mediana con relativo Std.Error, z-value e p-value.

Infine viene stimato il modello di Poisson che restituisce le stime dei parametri non distorte in media, allo scopo di ottenere un primo confronto delle tre diverse procedure. In riferimento al primo modello stimato tramite lo stimatore classico, quasi tutte le stime delle variabili risultano essere maggiori in valore assoluto, ad eccezione dell'intercetta, la quale sottolinea la presenza di un minor rischio di morte causato da neoplasia mammaria per il gruppo di riferimento. Per i pazienti appartenenti al genere maschile si assiste ad un leggero aumento del rischio di morte rispetto ai due modelli precedenti: infatti, si passa da un valore pari a 0.39 circa per i primi due modelli a un valore di 0.41. Nuovamente è possibile osservare una diminuzione del rischio di decesso per cancro al seno fino ai 50 anni, in relazione ai soggetti con un età compresa tra i 17 e 35 anni, seguito da un continuo aumento per i pazienti più anziani. Confrontando le stime ottenute con quelle non distorte in mediana non vi sono sostanziali differenze nella quasi totalità dei casi se non per l'intercetta e la variabile **Gender**, di cui si è già parlato. Anche in quest'ultimo caso la variabile inerente il genere dell'individuo non sembra essere utile nel prevedere il fenomeno di interesse.

Coefficiente	Stima	Std. Error	z-value	Pr(>|z|)
Intercetta	-0,37168	0,55724	-0,667	0,50477
etanew[35-40)	-0,46045	0,19464	-2,366	0,018001 *
etanew[40-45)	-0,75531	0,17224	-4,385	1,16e-05 ***
etanew[45-50)	-0,85457	0,15224	-5,365	8,09e-08 ***
etanew[50-55)	-0,69951	0,15416	-4,538	5,69e-06 ***
etanew[55-60)	-0,53537	0,15287	-3,502	0,000462 ***
etanew[60-65)	-0,41754	0,15207	-2,746	0,006037 **
etanew[65-70)	-0,32270	0,15157	-2,129	0,03325 *
etanew[70-75)	-0,18678	0,15488	-1,206	0,227843
etanew[75-80)	0,41812	0,15997	2,614	0,008956 **
etanew[80-85)	0,78483	0,17295	4,538	5,68e-06 ***
etanew[85-90)	1,19051	0,20706	5,75	8,95e-09 ***
GenderMale	0,41338	0,21235	1,947	0,051568
LocalizationCentral	-0,45586	0,1237	-3.685	0,000229 ***
LocalizationLateral	-0,51077	0,1181	-4,325	1,53e-05 ***
LocalizationMedian	-0,29773	0,12158	-2,449	0,01433 *
LocalizationWhole breast	0,09297	0,13733	0,677	0,4984
MetastasisLocalized	-1,70805	0,5325	-3,208	0,001338 **
MetastasisRegional	-1,2702	0,47603	-2,668	0,007624 **
StadiumII	0,39542	0,24766	1,597	0,110351
StadiumIII	0,74814	0,24681	3,031	0,002436 **
StadiumIV	0,12671	0,53586	0,236	0,813080
anno.entrata.intervallo[1970-1975)	-0,21001	0,04398	-4,775	1,80e-06 ***
anno.entrata.intervallo[1975-1980)	-0,55727	0,05488	-10,155	< 2e-16 ***
anno.entrata.intervallo[1980-1985)	-0,87261	0,06504	-13,417	< 2e-16 ***
anno.entrata.intervallo[1985-1990)	-1,13402	0,08095	-14,009	< 2e-16 ***

Tabella 4.6 – Stime dei parametri non distorti in media con relativo Std.Error, z-value e p-value .

CAPITOLO V

STUDI DI SIMULAZIONE

5.1 Svolgimento delle simulazioni

Nella parte finale della tesi sono stati svolti degli studi di simulazione Monte Carlo, con l'obiettivo di confrontare le proprietà dello stimatore non distorto in mediana con le proprietà dello stimatore "classico", ottenuto dalla funzione di verosimiglianza presentata al Capitolo 2. Inoltre, viene studiato anche lo stimatore non distorto in media, presentato da (Firth, 1993), come riferimento. Le simulazioni sono state eseguite tramite il software R, per differenti scelte della numerosità campionaria ($n = 100$, 150 e 200 osservazioni), della percentuale di censura (20% e 40% di casi censurati) e del numero di parametri di disturbo.

Il primo passo consiste nel generare in maniera casuale dalla distribuzione delle variabili esplicative un numero di osservazioni pari alla numerosità campionaria n scelta. Sia X_1 una variabile casuale con distribuzione Normale, $X_1 \sim N(0\,, 1.2)$ e X_2 una Bernoulliana, dove la probabilità di successo viene fissata pari a: $p = 0.6$. I valori dei coefficienti di regressione associati a queste due variabili sono fissati pari a $\beta_1 = 1.5$ e $\beta_2 = -1.5$. Successivamente viene generata la distribuzione della variabile tempo di censura che, in questo caso, è caratterizzata da una distribuzione Uniforme. Si ha quindi $C_i \sim U(0, a)$, dove a viene fissato in base alla proporzione di censura desiderata. Per ciascuna simulazione, si ipotizza che i tempi di attesa all'evento per ciascun soggetto, t_i, provengano da una distribuzione esponenziale, ossia $T_i \sim Exp(\lambda)$. Per generare da una distribuzione esponenziale occorre fissare un valore per il parametro λ, che rappresenta il tasso con la quale si verificano gli eventi, che per definizione sotto tale distribuzione è costante. È possibile costruire un modello di regressione esponenziale, il cui tasso è definito come:

$$rate = \lambda * \exp(x1 * \beta_1 + x2 * \beta_2),$$

dove viene fissato il tasso dell'evento per un soggetto con variabili esplicative nulle pari a $\lambda = 0.2$.

Successivamente viene considerata la variabile tempo osservato $\tilde{T}_i = min(C_i, T_i)$, le cui realizzazioni sono calcolate come il minimo tra il valore t_i e il valore della censura c_i, per ciascun soggetto i del campione, e rappresenta quindi il periodo di follow-up osservato. Si dispone inoltre di una variabile dicotomica, d_i, che informa se il soggetto i-esimo ha sperimentato l'evento di interesse o meno.

Il passo successivo riguarda la costruzione del dataset, la cui struttura deve essere del tutto identica a quella vista per i dati reali (Capitolo 4). Per ogni soggetto i, di ciascun dataset generato, si divide il relativo periodo di follow-up osservato in un dato numero di sotto-intervalli, andando così a generare diverse sotto-osservazioni per ogni individuo: per ogni soggetto sarà riportato il tempo di entrata e di uscita di ciascun intervallo e un indicatore che informerà se, nell'ultimo intervallo visitato, esso avrà sperimentato l'evento di interesse. La funzione utilizzata per la costruzione del dataset è la stessa di quella vista in precedenza, *TimeSplitter*, la quale permette di suddividere il periodo di follow-up di ciascun soggetto in un certo numero di sotto-intervalli, la cui ampiezza non deve essere necessariamente costante. La suddivisione in sotto-intervalli è la stessa per tutti i soggetti. Lo step finale consiste nel creare una variabile categoriale, chiamata **int.temporali**, le cui modalità rappresentano l'intervallo j visitato dal soggetto i-esimo. Questa verrà poi inserita nel modello e le rispettive classi fungeranno da parametri di disturbo. Il numero di parametri di disturbo è stato fissato pari a 5 in corrispondenza di una percentuale di censura del 20%, e pari a 4 per il 40% di censure.

Il numero R di replicazioni, per ciascuna simulazione, è stato fissato pari a 5000. Quindi, per ogni r-esima prova, vi sono questi due step di generazione delle variabili casuali e costruzione del dataset. La parte finale prevede l'applicazione dei tre differenti metodi di stima per il modello di Poisson adattato per l'analisi di sopravvivenza: il metodo classico (Capitolo 2), i metodi che restituiscono le stime dei parametri non distorte in mediana (Capitolo 4) e non distorte in media (Firth, 1993). I risultati sono le stime dei parametri e i relativi standard errors. Le stime sotto il modello di Poisson, ipotizzando lo stimatore non distorto in media di Firth, si ottiene specificando l'argomento **type = "As_mean"**, da inserire internamente al comando "glm".

Per verificare l'adeguatezza delle diverse procedure applicate nelle simulazioni e quindi l'appropriatezza del metodo che restituisce le stime dei parametri non distorte in mediana, che è di diretto interesse per la tesi, si procederà al confronto di appositi indici.

Tali indici sono:

- **PU** *(Probability of Underestimation)*: la frequenza percentuale con cui il parametro stimato è inferiore al vero valore del parametro;

- **MAE** *(Median Absolute Error)*: una misura di dispersione statistica definita come la mediana degli scarti, in valore assoluto, delle stime dei parametri dalla mediana (nota anche come *MAD*). In formule:

$$MAD = median\left(\left|\hat{\beta}_r - \hat{\beta}_{med}\right|\right), \qquad con\ \hat{\beta}_{med} = median\left(\hat{\beta}\right)$$

  dove $\hat{\beta}_r$ è la stima del parametro ottenuta nella replicazione r-esima e β il vettore delle stime dei coefficienti ottenute nelle varie replicazioni.

- **B** *(Bias)*. Usando le R replicazioni stimo il bias come:

$$E\left(\hat{\beta}\right) - \beta = \left(\frac{1}{R}\sum_{r=1}^{R}\hat{\beta}_r\right) - \beta = \frac{1}{R}\sum_{r=1}^{R}\left(\hat{\beta}_r - \beta\right),$$

  dove β rappresenta il vero valore del parametro;

- **RMSE** (*Root Mean Squared Error*): la radice quadrata dell'errore quadratico medio, il quale rappresenta una misura di accuratezza della stima ottenuta;

$$RMSE\left(\hat{\beta}\right) = \sqrt{MSE\left(\hat{\beta}\right)} = \sqrt{E\left[\left(\hat{\beta} - \beta\right)^2\right]},$$

utilizzando le R replicazioni, l'indice può essere riscritto come:

$$RMSE(\hat{\beta}) = \sqrt{\frac{1}{R}\sum_{r=1}^{R}\left[\left(\hat{\beta}_r - \beta\right)^2\right]}.$$

Viene inoltre fornita la probabilità di copertura al 95% per un intervallo di confidenza alla **Wald**, specificando la percentuale di volte in cui il valore vero del parametro cade all'interno, a destra dell'estremo superiore e a sinistra dell'estremo inferiore degli intervalli stimati per ciascuna replicazione.

5.2 Risultati e conclusioni

I risultati delle simulazioni svolte sono riportate nella Tabella 5.1. Le righe $\hat{\beta}_1$, $\tilde{\beta}_1$ e β_1^* riportano la stima "classica", non distorta in mediana e non distorta in media, rispettivamente, del parametro β_1. La stessa nomenclatura è stata applicata per le stime del parametro β_2. Gli stimatori vengono comparati tra loro tramite gli indici descritti in precedenza, i quali costituiscono le colonne della tabella; vengono inoltre riportati i risultati per i diversi valori della numerosità campionaria e della percentuale di censura. La PU è l'indice di primario interesse per valutare la bontà dello stimatore non distorto in mediana, dato che lo stesso è stato appositamente costruito per essere non distorto in mediana fino al terzo ordine. Secondo questo indice, è possibile osservare come la centratura in mediana per $\tilde{\beta}_1$ sia migliore rispetto alle altre due tipologie di stimatori, anche se vi è una differenza pressoché minima. Lo stesso, tuttavia, non si può dire per il parametro $\tilde{\beta}_2$, dato che la frequenza ottenuta con il metodo "classico" sembra avvicinarsi maggiormente al valore desiderato, ovvero al 50%: nella maggior parte dei casi $\tilde{\beta}_2$ sembra raggiungere una centratura migliore o quasi uguale a β_2^*.

Il parametro β_1^* presenta valori della distorsione che sono leggermente più bassi rispetto a quelli riportati dagli altri coefficienti, per entrambe le percentuali di

censura. Per β_2, invece, $\hat{\beta}_2$ sembra riportare una distorsione minore rispetto agli altri parametri nella maggior parte dei casi. Le differenze in termini di distorsione, si assottigliano all'aumentare della numerosità campionaria, tanto che per i campioni aventi dimensione pari a 200, i valori tendono a coincidere o quasi.

Prendendo in considerazione il MAE e l'RSME, si osserva come non vi sia quasi nessuna differenza tra i tre metodi utilizzati, tanto che al crescere di n i valori ottenuti sono praticamente gli stessi.

Infine, si osserva come in tutte le simulazioni svolte vi sia una maggiore probabilità di copertura per entrambi i parametri ottenuti col metodo "classico" (Capitolo 2).

Dai risultati delle simulazioni eseguite, si osserva come entrambi i metodi di correzione proposti diano risultati soddisfacenti. Nella maggior parte dei casi, essi presentano degli indici di distorsione e PU leggermente migliori rispetto a quelli ottenuti dallo stimatore "classico", raggiungendo di fatto l'obiettivo di riduzione della distorsione, anche se le differenze risultano essere minime, soprattutto per numerosità campionarie elevate. Infine i tre metodi sembrano presentare la medesima accuratezza, dato che gli indici relativi al MAE e al RSME, ottenuti dalle simulazioni, sono molto simili tra loro.

Si decide di svolgere delle simulazioni scegliendo una distribuzione esponenziale per la variabile censura. Le prove vengono eseguite per gli stessi valori della numerosità campionaria e percentuale di censura visti in precedenza. I risultati ottenuti sono disponibili nella Tabella 5.2 e di seguito analizzati. A differenza del caso precedente, la centratura in mediana per β_1^* risulta essere migliore rispetto a $\tilde{\beta}_1$, per qualsiasi valore di censura. Lo stesso non si può dire per β_2^*, dato che, per una percentuale di censura del 20%, viene raggiunta una centratura leggermente migliore dallo stimatore classico $\hat{\beta}_2$, anche se questa minima differenza tende a calare all'aumentare della numerosità campionaria. Per il 40% di censura, invece, si assiste a una situazione inversa.

Come nel caso precedente, il metodo di Firth restituisce una minor distorsione per il parametro β_1, mentre nel caso del parametro β_2 il metodo classico risulta essere

quello preferibile, anche se queste differenze sono minime e pressoché inesistenti al crescere di n. Nuovamente gli indici inerenti il MAE e l'RSME sembrano essere equivalenti.

Infine, in tutti i risultati ottenuti, è presente una probabilità di copertura maggiore per gli stimatori $\hat{\beta}_1$ e $\hat{\beta}_2$, anche se vi è una differenza minima.

Confrontando i risultati appena riportati con quelli ottenuti nel caso di una censura uniforme, si osserva come i due metodi di correzione applicati sembrano ridurre la distorsione in tutti i casi, anche se non si assistono a sostanziali differenze con il metodo "classico".

Param.	n	Vero val. par.	Censura	PU	MAE	B	RSME	Wald	Coda sin. Wald	Coda dx. Wald
$\hat{\beta}_1$	100	1,5	20%	44,6	0,1462	0,0277	0,1515	94,92	3,52	1,56
$\tilde{\beta}_1$	100	1,5	20%	44,92	0,146	0,0271	0,1513	94,66	3,68	1,66
β_1^*	100	1,5	20%	44,9	0,146	0,0266	0,1512	94,5	3,74	1,76
$\hat{\beta}_2$	100	-1,5	20%	54,68	0,259	-0,0323	0,2583	94,74	2.08	3,18
$\tilde{\beta}_2$	100	-1,5	20%	54,76	0,259	-0,0331	0,2583	94,3	2.16	3,54
β_2^*	100	-1,5	20%	54,76	0,2583	-0,0334	0,2578	94,18	2.22	3,6
$\hat{\beta}_1$	150	1,5	20%	44,94	0,1187	0,0188	0,1192	94,72	3,5	1,78
$\tilde{\beta}_1$	150	1,5	20%	45	0,1188	0,0184	0,1191	94,54	3,56	1.90
β_1^*	150	1,5	20%	44,98	0,1189	0,0183	0,1192	94,32	3,72	1,96
$\hat{\beta}_2$	150	-1,5	20%	53,54	0,2029	-0,0188	0,2068	94,16	2,46	3,38
$\tilde{\beta}_2$	150	-1,5	20%	53,66	0,2029	-0,0196	0,2069	93,92	2,58	3,5
β_2^*	150	-1,5	20%	53,84	0,2024	-0,0202	0,2066	93,86	2,6	3,54
$\hat{\beta}_1$	200	1,5	20%	46,1	0,0995	0,0132	0,1024	94,42	3,5	2,08
$\tilde{\beta}_1$	200	1,5	20%	46,16	0,0996	0,0129	0,1024	94,22	3,6	2,18
β_1^*	200	1,5	20%	46,14	0,0999	0,0129	0,1024	94,12	3,64	2,24
$\hat{\beta}_2$	200	-1,5	20%	52,76	0,1723	-0,0143	0,1715	95,08	1,72	3,2
$\tilde{\beta}_2$	200	-1,5	20%	52,82	0,1714	-0,0149	0,1715	94,94	1,76	3,3
β_2^*	200	-1,5	20%	52,92	0,1711	-0,0153	0,1714	94,84	1,82	3,34
$\hat{\beta}_1$	100	1,5	40%	44,92	0,1539	0,0277	0,1604	94,58	3,78	1,64
$\tilde{\beta}_1$	100	1,5	40%	45,08	0,1537	0,0271	0,1603	94,4	3,86	1,74
β_1^*	100	1,5	40%	45,08	0,1539	0,0266	0,1602	94,12	4,02	1,86
$\hat{\beta}_2$	100	-1,5	40%	53,2	0,2733	-0,0285	0,2756	94,66	2,16	3,18
$\tilde{\beta}_2$	100	-1,5	40%	53,16	0,273	-0,0289	0,2754	94,54	2,18	3,28
β_2^*	100	-1,5	40%	53,12	0,2722	-0,0285	0,2756	94,3	2,3	3,4
$\hat{\beta}_1$	150	1,5	40%	45,4	0,1308	0,0195	0,1307	94,68	3,48	1,84
$\tilde{\beta}_1$	150	1,5	40%	45,5	0,1312	0,0193	0,1307	94,52	3,58	1,9
β_1^*	150	1,5	40%	45,42	0,1309	0,0192	0,1307	94,38	3,72	1,9
$\hat{\beta}_2$	150	-1,5	40%	52,5	0,23	-0,0185	0,234	93,8	2,68	3,52
$\tilde{\beta}_2$	150	-1,5	40%	52,54	0,2293	-0,0188	0,2338	93,68	2,72	3,6
β_2^*	150	-1,5	40%	52,54	0,2291	-0,0184	0,2332	93,6	2,78	3,62
$\hat{\beta}_1$	200	1,5	40%	46,8	0,1056	0,0136	0,1101	94,24	3,5	2,26
$\tilde{\beta}_1$	200	1,5	40%	46,88	0,1055	0,0134	0,1101	94,2	3,54	2,26
β_1^*	200	1,5	40%	46,8	0,1055	0,0134	0,1102	94,06	3,6	2,34
$\hat{\beta}_2$	200	-1,5	40%	51,2	0,1866	-0,0111	0,187	95,12	1,96	2,92
$\tilde{\beta}_2$	200	-1,5	40%	51,16	0,186	-0,0115	0,187	95	2	3
β_2^*	200	-1,5	40%	51,06	0,1855	-0,0115	0,1867	94,86	2,1	3,04

Tabella 5.1 – Risultati delle simulazioni con censura uniforme.

Param.	n	Vero val. par.	Censura	PU	MAE	B	RSME	Wald	Coda sin. Wald	Coda dx. Wald
$\hat{\beta}_1$	100	1,5	20%	42,78	0,1415	0,0309	0,1498	94,7	3,6	1,7
$\tilde{\beta}_1$	100	1,5	20%	42,92	0,1415	0,0302	0,1496	94,52	3,74	1,74
β_1^*	100	1,5	20%	43,08	0,1415	0,0297	0,1496	94,32	3,84	1,84
$\hat{\beta}_2$	100	-1,5	20%	53,46	0,2444	-0,0315	0,2515	94,68	1,76	3,56
$\tilde{\beta}_2$	100	-1,5	20%	53,52	0,2433	-0,0324	0,2516	94,36	1,8	3,84
β_2^*	100	-1,5	20%	53,6	0,2424	-0,0329	0,2511	94,18	1,82	4
$\hat{\beta}_1$	150	1,5	20%	44,38	0,1175	0,0197	0,1208	94,38	3,52	2,1
$\tilde{\beta}_1$	150	1,5	20%	44,52	0,1176	0,0194	0,1207	94,22	3,6	2,18
β_1^*	150	1,5	20%	44,6	0,1181	0,0192	0,1207	94,08	3,66	2,26
$\hat{\beta}_2$	150	-1,5	20%	53,7	0,1954	-0,0201	0,2022	94,62	2,5	2,88
$\tilde{\beta}_2$	150	-1,5	20%	53,88	0,1949	-0,0208	0,2022	94,46	2,56	2,98
β_2^*	150	-1,5	20%	53,86	0,1945	-0,0212	0,2020	94,22	2,64	3,14
$\hat{\beta}_1$	200	1,5	20%	45.84	0,1018	0,0143	0,1027	94,76	3,24	2
$\tilde{\beta}_1$	200	1,5	20%	45.98	0,1018	0,0141	0,1027	94,62	3,32	2,06
β_1^*	200	1,5	20%	45.94	0,1019	0,0141	0,1027	94,48	3,4	2,12
$\hat{\beta}_2$	200	-1,5	20%	52,42	0,1712	-0,0146	0,1728	94,98	2,12	2,9
$\tilde{\beta}_2$	200	-1,5	20%	52,54	0,1712	-0,0153	0,1728	94,8	2,18	3,02
β_2^*	200	-1,5	20%	52,56	0,1715	-0,0157	0,1727	94,74	2,22	3,04
$\hat{\beta}_1$	100	1,5	40%	43,12	0,1534	0,0334	0,1647	94,62	3,68	1,7
$\tilde{\beta}_1$	100	1,5	40%	43,40	0,1530	0,0328	0,1646	94,48	3,78	1,74
β_1^*	100	1,5	40%	43,56	0,1532	0,0322	0,1646	94,3	3,94	1,76
$\hat{\beta}_2$	100	-1,5	40%	53,64	0,2801	-0,0336	0,2853	94,8	1,92	3,28
$\tilde{\beta}_2$	100	-1,5	40%	53,58	0,2816	-0,0339	0,2850	94,44	2,04	3,52
β_2^*	100	-1,5	40%	53,34	0,2809	-0,0332	0,2840	94,24	2,14	3,62
$\hat{\beta}_1$	150	1,5	40%	44,98	0,1244	0,0199	0,1321	94,32	3,42	2,26
$\tilde{\beta}_1$	150	1,5	40%	45.04	0,1242	0,0196	0,132	94,22	3,5	2,28
β_1^*	150	1,5	40%	45,12	0,1239	0,0194	0,1321	94,02	3,64	2,34
$\hat{\beta}_2$	150	-1,5	40%	52,9	0,2249	-0,0167	0,227	94,7	2,42	2,88
$\tilde{\beta}_2$	150	-1,5	40%	53	0,2256	-0,0170	0,2268	94,6	2,44	2,96
β_2^*	150	-1,5	40%	52,88	0,2251	-0,0167	0,2263	94,5	2,54	2,96
$\hat{\beta}_1$	200	1,5	40%	45,36	0,1108	0,0161	0,1124	94,9	3,26	1,84
$\tilde{\beta}_1$	200	1,5	40%	45,3	0,1110	0,016	0,1123	94,82	3,32	1,86
β_1^*	200	1,5	40%	45,32	0,1108	0,016	0,1124	94,64	3,44	1,92
$\hat{\beta}_2$	200	-1,5	40%	52,32	0,1955	-0,0143	0,1932	94,94	2,28	2,78
$\tilde{\beta}_2$	200	-1,5	40%	52,36	0,1951	-0,0146	0,1931	94,78	2,3	2,92
β_2^*	200	-1,5	40%	52,3	0,1948	-0,0146	0,1928	94,6	2,34	3,06

Tabella 5.2 – Risultati delle simulazioni con censura esponenziale.

CAPITOLO VI

CONCLUSIONI

6.1 Conclusioni generali

Lo scopo di questo lavoro consiste nella presentazione del modello di Poisson adattato in un contesto di analisi di sopravvivenza, un campo che differisce da quello che è il suo normale utilizzo. Per conferire un valore aggiunto alla tesi, per il modello studiato, si è deciso di confrontare il relativo stimatore di massima verosimiglianza con lo stimatore non distorto in mediana (Kenne Pagui, Salvan, & Sartori, 2017) e con lo stimatore non distorto in media di Firth.

In primis è stato dimostrato come la funzione di verosimiglianza per osservazioni censurate a destra di un modello esponenziale a rischi proporzionali e costanti a tratti risulti essere del tutto equivalente a quella del modello di Poisson, nell'ipotesi in cui si abbia a disposizione un periodo di esposizione t_{ij} per ciascun soggetto i, per i diversi intervalli j che lo stesso visita, e una variabile risposta $d_{ij} \sim Poisson(\mu_{ij})$, la quale rappresenta un indicatore di evento o censura. Con l'intento di dare una dimostrazione pratica dell'uso di questo modello, si è deciso di procedere alla sua applicazione in un contesto reale, in cui si studia la sopravvivenza di un gruppo di pazienti norvegesi affetti da cancro al seno. L'età è stata presa come scala temporale di riferimento. I risultati hanno dimostrato come vi sia un differente rischio di decesso per le varie fasce d'età dei soggetti, giustificato dal fatto che i coefficienti associati alle variabili inerenti gli intervalli d'età siano risultati quasi tutti significativi. Prendendo come riferimento i soggetti più giovani dello studio, il rischio di morte associato alla neoplasia mammaria sembra avere un andamento calante fino ai pazienti aventi un'età compresa tra i 45-50 anni, per poi aumentare considerevolmente nei soggetti più anziani. La maggior parte delle caratteristiche socio-demografiche degli individui sono risultate essere utili nel prevedere e interpretare il fenomeno di interesse, fatta a eccezione per la variabile rappresentante il genere. Confrontando successivamente le stime di massima verosimiglianza con quelle ottenute dagli stimatori non distorti in mediana e in media, si è osservato come vi siano leggere differenze

nelle stime dei parametri, anche se l'interpretazione del fenomeno rimane comunque la stessa in tutti e tre i casi.

L'obiettivo secondario della tesi risiede nel mettere a confronto le proprietà delle tre diverse tipologie di stimatori. Nel Capitolo V vengono presentati degli studi di simulazione Monte Carlo per diverse scelte della numerosità campionaria e percentuale di censura, i quali sono stati svolti proprio a tal fine. Nel caso in cui venga fissata una distribuzione Uniforme per la variabile censura, si osserva come entrambi i metodi di correzione proposti diano risultati soddisfacenti nel ridurre la distorsione, anche se le differenze individuate sono minime. Infatti, a differenza dello stimatore "classico", gli stimatori proposti da (Kenne Pagui, Salvan, & Sartori, 2017) e da (Firth, 1993) presentano una centratura in mediana e una distorsione di poco migliori, mentre osservando gli indici inerenti l'accuratezza dello stimatore si evince come i tre metodi analizzati diano risultati molto simili tra loro. Nella seconda parte delle simulazioni si decide di assegnare alla variabile censura una distribuzione Esponenziale, con lo scopo di verificare se l'ipotesi di una distribuzione uniforme per la variabile censura avesse inciso sugli esiti della procedura. Nuovamente non sembrano esserci differenze decisive nel preferire una tipologia di stimatore rispetto ad un altro, poiché gli indici ottenuti non sembrano riportare sostanziali differenze, se non una leggera centratura in mediana migliore per lo stimatore di Firth. Anche in questo secondo caso le differenze rispetto alle altre due tipologie di stimatori non sono risultate essere sostanziali. Analizzando i risultati ottenuti nel loro insieme, sembra che i due metodi di correzione proposti riescano nel tentativo di ridurre la distorsione, anche se le differenze che si osservano con lo stimatore "classico" sono minime, soprattutto in termini di accuratezza.

6.2 Possibili estensioni ad altri modelli

In questo paragrafo viene approfondita la possibilità di costruire un modello alternativo, basato sulla: *Poisson-Weibull distribution* (*PW*), il cui campo di applicazione principale è lo stesso del modello di Poisson discusso in precedenza: l'analisi della sopravvivenza. Di seguito verrà riportata una breve presentazione del modello, i cui fondamenti teorici fanno riferimento al lavoro condotto da (Bereta, Louzada, & Franco, 2011).

Si consideri un campione casuale di grandezza $N \geq 1$ generato da una variabile casuale Y, avente distribuzione $Weib(\beta, \gamma)$, con $\beta > 0$ parametro di scala e $\gamma > 0$ parametro di forma. La relativa funzione di densità f_0 è data da:

$$f_0(y;\, \beta, \gamma) = \gamma \beta^{\gamma} y^{\gamma-1} e^{-(\beta y)^{\gamma}} I_{[0,\infty)}(y), \tag{6.1}$$

dove $I_{[0,\infty)}(y)$ è la funzione indicatrice rappresentante l'insieme dei valori che può assumere la distribuzione Weibull.

Sia N una variabile casuale discreta caratterizzata da una distribuzione Poisson troncata a zero, di parametro $\alpha > 0$. Allora, la funzione di densità di probabilità f_1 sarà pari a:

$$f_1(n; \alpha) = \frac{\alpha^n}{n!(e^{\alpha}-1)} I_{\{1,2,\dots\}}(n), \tag{6.2}$$

dove $I_{\{1,2,\dots\}}(n)$ è la funzione indicatrice rappresentante l'insieme dei valori che può assumere la distribuzione Poisson.

La funzione di densità del Poisson-Weibull si ottiene dalla mistura delle due funzioni di densità definite nella (6.1) e (6.2).

Sia T una variabile casuale non negativa la cui funzione di densità si ottiene come valore minimo di N distribuzioni Weibull, ciascuna generata da una variabile casuale i.i.d. $Y_i \sim Weib\left(N^{\frac{1}{\gamma}}\beta, \gamma\right)$, con N variabile avente distribuzione Poisson definita nella (6.2). Si ottiene così la densità per la *PW*, la cui dimostrazione è riportata in (Bereta, Louzada, & Franco, 2011):

$$f_T(t;\alpha,\beta,\gamma) = \frac{e^{\alpha\, e^{-(\beta t)^\gamma} - (\beta t)^\gamma}\, \alpha\beta^\gamma\, t^{\gamma-1}\gamma}{e^\alpha - 1}\, I_{[0,\infty)}(t), \quad con\ \alpha,\beta,\gamma > 0 \quad (6.3)$$

Vengono presentate di seguito la relativa funzione di ripartizione F_T e la funzione di rischio h_T, la quale, essendo specificata da tre parametri, risulta abbastanza flessibile e può difatti assumere diversi andamenti (crescente, decrescente o "bathtub") a seconda dei valori scelti peri i coefficienti:

$$F_T(t;\alpha,\beta,\gamma) = \frac{(e^\alpha - e^{\alpha\, e^{-(\beta t)^\gamma}})}{e^\alpha - 1}\, I_{[0,\infty)}(t), \quad con\ \alpha,\beta,\gamma > 0 \quad (6.4)$$

$$h_T(t;\alpha,\beta,\gamma) = \frac{e^{\alpha\, e^{-(\beta t)^\gamma} - (\beta t)^\gamma}\, \gamma\alpha\beta^\gamma t^{\gamma-1}}{e^{\alpha\, e^{-(\beta t)^\gamma}} - 1}\, I_{[0,\infty)}(t), \quad con\ \alpha,\beta,\gamma > 0 \quad (6.5)$$

Assumendo un campione di numerosità n, generato casualmente da una distribuzione PW, avente la medesima funzione di densità specificata dalla (6.3), è possibile ottenere la stima di massima verosimiglianza del vettore di parametri $\theta = (\alpha, \beta, \gamma)^T$, massimizzando la relativa funzione di log-verosimiglianza:

$$l(\theta;t) = \alpha \sum_{i=1}^{n} e^{-(\beta t_i)^\gamma} - \sum_{i=1}^{n} (\beta t_i)^\gamma + n \log\alpha + n\gamma \log\beta +$$

$$+\gamma\sum_{i=1}^{n} logt_i - \sum_{i=1}^{n}\log t_i + n\log\gamma - n\log(e^{\alpha}-1) \qquad (6.6)$$

Sotto condizioni di regolarità (Pace & Salvan, 2001), la distribuzione asintotica del vettore dei parametri θ è:

$$\sqrt{n}\left(\hat{\theta} - \theta\right) \sim N_3(0,\, I(\theta)^{-1}), \qquad (6.7)$$

con $I(\theta)$ matrice di informazione attesa, di dimensione $(3x3)$, dalla quale è possibile calcolare le stime degli standard errors dei parametri. In alternativa, è possibile utilizzare l'inversa della matrice di informazione osservata, in quanto rappresenta uno stimatore consistente per $I(\theta)^{-1}$ (Pace & Salvan, 2001). Tale distribuzione asintotica può essere utilizzata per la costruzione di intervalli e regioni di confidenza di livello approssimato per i singoli parametri.

La distribuzione Poisson-Weibull rappresenta un nuovo ed efficace strumento, il cui campo applicativo riguarda principalmente l'analisi della sopravvivenza. Simulazioni svolte nel corso dello studio degli autori hanno dimostrato che l'inclusione di tre parametri nella funzione di densità del modello conducono a una funzione di rischio più flessibile rispetto alla distribuzione Weibull, e per questo motivo, spesso, ne si predilige l'utilizzo, anche se fino ad ora non ha trovato larga applicazione. Una proposta di notevole interesse sarebbe quella di costruire un modello di regressione per dati di sopravvivenza a partire dalla distribuzione Poisson-Weibull, con tasso costante a tratti.

I tre metodi di stima studiati nella tesi potrebbero essere usati anche sotto il modello più esteso Poisson-Weibull, ed i risultati potrebbero essere confrontati con quelli ottenuti sotto un modello di regressione di Poisson.

BIBLIOGRAFIA E SITOGRAFIA

Agresti, A. (2003). *Categorical Data Analysis.* New York: Wiley.

Allison, P. (1995). *Survival Analysis Using SAS: A Practical Guide.* Cary, Nc, USA: SAS Institute Inc.

Azzalini, A. (2001). Inferenza statistica: Una presentazione basata sul concetto di verosimiglianza. Springer Science & Business Media.

Barndorff-Nielsen. (1986). Inference on full or partial parameters based on the standardized signed log likelihood ratio. *Biometrika 73*, 307-322.

Barndorff-Nielsen, O. E., & Cox, D. R. (1989). Inference on full or partial parameters based on the standardized signed log likelihood ratio. *Biometrika 73*, 307–322.

Bereta, E. M., Louzada, F., & Franco, M. A. (2011). The Poisson-Weibull distribution. *Advances and Applications in Statistics 22*, 107-118.

Biehler, M., Holling, H., & Doeber, P. (2015). Saddlepoint approximations of the distribution of the person parameter in the two parameter logistic model. *Psychometrika 80*, 665-688.

Blossfeld, H., Golsch, K., & Rohwer, G. (2012). *Event History Analysis With Stata.* Psychology Press.

Cain, K., Harlow, S. D., Little, R. J., Nan, B., Yosef, M., Taffe, J. R., & Elliot, M. R. (2011). Bias Due to Left Truncation and Left Censoring in Longitudinal Studies of Developmental and Disease Processes. *American Journal of Epidemiology*, 1078-1084.

Cancer Registry of Norway. (2018, 09 7). Tratto da https://www.kreftregisteret.no/en/ Ultimo accesso: 04/12/2018

Cox, D. R. (1972). Regression Models and Life-Tables.

Dalgaard, P. (2008). *Introductory Statistics with R.* Springer Science & Business Media.

Firth. (1993). Bias reduction of maximum likelihood estimates. *Biometrika 80,* 27-38.

Guimmole, F., & Ventura, L. (2002). Practical point estimation from higher-order pivots. *J. Statist. Comp. Simul. 72*, 419-430.

Hirji, K. F., Tsiatis, A. A., & Mehta, C. R. (1989). Median unbiased estimation for binary data. *Am. Statistician 43*, 7-11.

Holford, T. (1980). *The Analysis of Rates and of Survivorship Using Log-Linear Models* . International Biometric Society.

Kenne Pagui, E. C., Salvan, A., & Sartori, N. (2017). Median bias reduction of maximum likelihood estimates . *Biometrika, Volume 104, Issue 4*, 923–938.

Kosmidis, I. (2014). Bias in parametric estimation: Reduction and useful side-effects. *WIREs Comp. Statist. 6*, 185-196.

Laird, N., & Olivier, D. (1981). *Covariance Analysis of Censored Survival Data Using Log-Linear Analysis Techniques.* Journal of the American Statistical Association.

Lehmann, E. L., & Romano, J. P. (2005). *Testing Statistical Hypotheses.* Springer.

McCullagh, P., & Tibshirani, R. J. (1990). A simple method for the adjustment of profile likelihoods. *J. R. Statist. Soc. B 52*, 325-344.

Oja, H. (2013). *OJA, H. (2013). Multivariate median. In Robustness and Complex Data Structures, C. Becker, R. Fried & S. Kuhnt.* Berlin: Springer.

Pace, L., & Salvan, A. (1997). *Principles of Statistical Inference from a Neo-Fisherian Perspective.* Singapore: World Scientic Pub Co Inc.

Pace, L., & Salvan, A. (1999). Point estimation based on confidence intervals: Exponential families. *J. Statist. Comp. Simul. 64*, 1-21.

Pace, L., & Salvan, A. (2001). *Introduzione alla statistica II. Inferenza, verosimiglianza, modelli.* Padova: Cedam.

Rodriguez, G. (2007). *Lecture Notes on Generalized Linear Models.* Princeton University.

Seeber, G. U. (2014). *Poisson Regression: Biostatistical Applications.* Wiley StatRef: Statistic Reference Online.

Stern, S. E. (1997). A second-order adjustment to the profile likelihood in the case of a multidimensional parameter of interest. *J. R. Statist. Soc. B 59*, 653-665.

World Health Organization. (2018, 09 7). Tratto da https://www.who.int/cancer/prevention/diagnosis-screening/breast-cancer/en/ Ultimo accesso: 04/12/2018

RINGRAZIAMENTI

Sono della convinzione che, nella vita, nessun traguardo importante possa essere raggiunto da soli, nonostante l'impegno, la serietà e la costanza che una qualsiasi persona possa metterci. Per questo motivo, desidero dedicare un piccolo spazio a tutte quelle persone che, con il loro affetto o con la forza delle loro parole, mi hanno consentito di arrivare a quello che è uno dei risultati, per me, più rilevanti e belli di sempre.

In primis vorrei ringraziare la mia famiglia, in particolare i miei genitori. Grazie per essermi sempre stati accanto qualunque fossero state la mie scelte, per avermi educato, per avermi sostenuto economicamente permettendomi di costruirmi un futuro e per avermi sempre voluto bene. Siete e rimarrete sempre un esempio che ciascun uomo dovrebbe seguire nella propria vita.

Desidero poi ringraziare la mia relatrice, la professoressa Cortese, per l'enorme disponibilità e gentilezza con la quale mi ha seguito nella realizzazione di questa tesi e il dottor Kenne Pagui per il suo prezioso contributo.

Un grazie anche ai mei compagni di corso, soprattutto ad Andrea, Sara e Maria, con cui ho condiviso gioie e sofferenze in questi due anni di studi. Non so cosa il futuro ci riservi, ma spero con tutto il cuore che farete ancora parte della mia vita.

Vorrei ringraziare anche gli amici, miei compagni di vita. Se oggi potete chiamarmi per la seconda volta dottore è anche merito del vostro continuo sostegno e vicinanza. Grazie soprattutto a voi, ex compagni della quinta F, che anche dopo anni continuate a rendermi felice e a volermi bene, anche se a modo vostro. Ma lo adoro.

Infine un ringraziamento speciale va a te. Il tuo amore, il tuo sostegno e le tue parole mi hanno sempre spinto a fare del mio meglio, in tutte le cose, e se oggi posso festeggiare quello che, per ora, è il traguardo più grande della mia vita lo devo soprattutto a te. Grazie Laura, per tutto quello che è stato e per quello che sarà.

Grazie a tutti, davvero.